info@WDP

From: Jared Hirst <jared.hirst@serversaustralia.com.au>
Sent: Friday, 8 September 2017 3:58 PM
To: info@wdpvoip.net.au
Subject: Post Incident Report - Sydney Outage 4/9/2017

ServersAustralia

Dear Frank

On the 4th of September 2017, at 1:00am Australian Eastern Standard Time, Servers Australia suffered from multiple network incidents that affected multiple sites within the Sydney network. The following Data Centre's were affected by packet loss and or complete loss of service for various lengths of time.

- Equinix Sydney 1
- Equinix Sydney 3
- Equinix Sydney 4
- Vocus Doody Street
- Syncom St Leonards
- SAU DC Wyong

Customers within the above locations saw packet loss and outages for durations from 5 minutes all the way up to 8 hours.

Over the past 4 days, I have been working with my team to ascertain what has gone wrong, how 6 independent sites could be taken offline all in one go, and how a small issue can cascade so quickly into a major issue, affecting all primary and secondary switches / routers and transit providers.

As most of our clients have seen, we have been investing heavily into our network in the past 3 years, and this is the first major outage that Servers Australia has had in a very long time. Outages are painful, and not something that should happen. I am very embarrassed that we have had an outage, especially after we have been upgrading and spending money to ensure that they don't happen, though outages can and do happen to everyone at some point.

There is no excuse or fixing what happened on the 4th, but I want to be as transparent as I can with our customers, therefore we have provided a complete detailed timeline of the exact events that happened at the bottom of this email, along with an action to ensure this does not happen again.

I have summarised the events, as the timeline is very detailed.

At 1:00 am on the 4th, our team noticed a few servers starting to see loss. The 24/7 support team investigated this and escalated to our Network team. From here, we had a string of issues that stemmed from a broadcast storm due to older core network equipment that was scheduled to be removed as part of our upgrades, there was compatibility issues between the new hardware and the old hardware, and as such, this caused a range of stacked and redundant switches to reboot themselves due to high CPU load, thus causing people to go offline.

We also had our remote access provider offline, which caused huge delays in getting access to our core equipment to be able to diagnose the issues, and there was also failing optics in a switch that were working fine, however, when they were under high traffic load, started to produce packet loss.

All these issues cascaded and got bigger and bigger as time went on. In total we had 9 different issues all happen within the 8 hour window. A few could have been avoided from our side, and we are going to work to fix those urgently, however, there were issues outside our control that we are going to work with the vendors to ensure they cannot happen in the future.

The majority of the issues do stem from the switch stacks overloading and rebooting, and the OSPF sessions all being in one area and not seeing the network as 'down'. Therefore, the redundant sites would come online for a few minutes and then back offline again.

We understand and acknowledge that there was also some issue with getting in contact with us, and I would like to ensure everyone knows that https://status.mysau.com.au will always be available, as this is 100% hosted off-net.

Our phone system was available during the outage, however due to the huge volume of calls, there was congestion, and some delays in putting up a status message. This has now been rectified with the system being moved to a larger server and a larger link that can handle well over 300+ concurrent calls, which is an increase of 5 times the previous limit.

I have included below, a detailed Root Cause Analysis of the events and the actions we are taking, below that is also a full timeline of events.

Root Cause Analysis (RCA)

Servers Australia have analysed this outage in detail and how we can prevent this from recurring. The following details how we will be resolving all issues moving forward to deliver a more resilient network:

1. Issue: Switches rebooting in SY4 without reason.
The switch stacks were rebooting in SY4 and SY1 in the wrong order and for no reason, these reboots caused significant downtime and are stacked to allow for redundancy, this caused a huge amount of downtime and is the number one priority to be rectified.
Action plan: Follow up with Switching Vendor SY4 – High Priority - 14 days.
A full review will take place with the switching vendor used in SY4 to analyse why the stack of switches locked and why they reloaded in the incorrect order. Remediation action will take place based on replication testing by the switching vendor

2. Issue: Legacy Equipment.
Servers Australia have gone through extremely rapid growth throughout the last 2 years leading to pockets within the network where legacy core network equipment remains. These devices were being replaced by new equipment and services moved over as the opportunity presents itself. To date 90% of the services connected to the core network have been moved to the new core with only 10% of services once connected to the old core remaining.
Action Plan: Expediate hardware replacement - High priority – 90 days.
Within the next 90 days, Servers Australia will be retiring all legacy core networking equipment across all Sydney locations. We expect minimal impact from these works taking place after hours,

the current plan was to retire this over 12 months, however the impact that this has caused on Monday has moved this forward.

3. Issue: Location of BGP "Pull up" routes.
Border Gateway Protocol (BGP) is the network protocol used to connect the Servers Australia network to rest of the internet. BGP pull up routes are special routes used to fully establish our BGP sessions to our Transit and Peering providers and mark us as online and ready to receive traffic from the internet. This role is currently completed by the legacy core network and as such when this device was taken offline, subsets of our core network routing were unavailable. Some routes remained online while others were withdrawn completely from the internet resulting in packet loss and services to drop offline in multiple locations.
Action Plan: BGP Pull up route relocation – High Priority – 14 days.
Over the next 14 days BGP pullup routes will be moved from the legacy core, to our border routers and switching hardware. These will also be made more redundant in this process.

4. Issue: Location of "Top Of Rack" Distribution network.
Top Of Rack is the term given to switches in each customer rack linking back to the core. A small set of services were still terminating onto this old core network resulting in a total outage while the switch was restored.
Action Plan: Location of Top Of Rack distribution network – 90 days.
Servers Australia will be accelerating the timeline of the Top of Rack uplink migrations forward. This will see all of our Top Of Rack switches migrated from the old Core network to our new core network. There will be a 2-3 minute disruption to customers during these works, we will work with customers during this time and ensure there is minimal impact.

5. Issue: Location of Servers Australia Management Services.
The location of the management services used by the Servers Australia corporate network were still directly connected to the Old Cisco Core network. This included services like network configuration backups. These are accessible over both the Out of Band network and the production network to allow us to have access at any time. When the Cisco Core Network went offline, we lost access to the production network and went to use the Out of Band network which was also offline leaving us in a situation where we were unable to access these at all to utilise them and forcing us to enact part of our Disaster Recovery (DR) Plan involving our fail safe configurations and in turn leading to a longer and more complex Time To Resolution (TTR).

Action Plan (60 Days): Network Engineering, Infrastructure and Support will be working closely to move our management systems into multiple locations. We will also be working to keep a tertiary copy of this critical information within our office as well as within the core networks that will comply with our ISO9001 operating procedures.

6. Issue: Unavailability of our Out of Band network.
Servers Australia operate an Out of Band network for all network equipment in all datacentres connected via either 4G or ADSL connections (depending on availability at each datacentre). These connections are completely separate to our network to ensure that we can access them at any time no matter what is occurring within our network Our SY1 Out of Band Network is connected via ADSL. During the outage we experienced, our ADSL provider had an outage on the connection lasting more than 4 hours exceeding the SLA we hold. This in turn lead us to use Smart Hands Services to provide Out Of Band access leading to further delays. The chance of our OOB provider being offline at the same time that we were offline was incredibly rare and very unlucky. This now has us reviewing all our OOB systems, and is detailed in the Action plan.
Action Plan (60 Days): Whilst we utilise a business grade off-net connection for Out of Band, it was offline too long and created significant delays in rectification. Servers Australia will be completing a redesign of our Out of Band networks which will include multiple diverse providers using different technology access methods.

7. Issue: OSPF Routing Inefficiencies.
Open Shortest Path First (OSPF) is the routing protocol used within the Servers Australia network to distribute IP blocks and connectivity around the whole network. When links rapidly transition between Up and Down states (flapping) like we saw in this outage, it causes a reconvergence of parts of the OSPF network resulting in packet loss across the network while this takes place
Action Plan (60 Days): Servers Australia will be completing a review of the OSPF layout and areas with our Network Engineers working with external experts to ensure the work we are completing conforms with industry best practices. Whilst our staff are all extremely qualified, as the network has grown, we have grown with it. As peace of mind for our customers we are having our plans independently audited to review the layout and suggest improvements.

8. Issue: Delays in configuration Merge back from Failsafe configuration.
As shown above there were significant delays in the merge of the backup configuration into the failsafe configuration. In our DR testing this process took less than 5 minutes and was considered an acceptable TTR to be incorporated into the plan. With the high CPU load caused by the

broadcast storm, this merge took far longer than expected and anticipated and hindered our ability to resolve the root of the issues. Additionally, the merge that was completed did not fully replace the failsafe configuration and redundant configuration was activated that should not have been.
Action Plan (30 Days): The restoration and merge from failsafe configurations was expected to take around 5 minutes. In light of this issue, we will be amending our DR policies and changing how this part is completed. This will make us far more resilient to having to utilise these configurations at all.

9. Issue: Broadcast storm and temporary fix.
The broadcast storm was caused by a switch unwrapping a QinQ VLAN and causing a huge storm between two VLANS that had traffic appearing in two areas at the same time, this then caused a spanning tree issue, and subsequently caused a broadcast storm between our old and new network with older VLANs. The fix that was put in place was reliable enough at the time of this issue occurring at around 1:30am, however later on in the morning when the configuration merge happened, it activated a piece of configuration causing a loop and thus another broadcast storm. Once the merge had completed this was able to be rectified.
Action Plan (90 Days): Legacy VLANs will be reviewed and shutdown where possible moving from spanned VLANs to Routed Ports injected to OSPF. In doing so we remove the ability for broadcast storms to exist and removing the need for storm-control on our switches.

All of the above will not be easy to do in a short time, therefore we are recruiting more Network Engineering staff immediately to ensure that we can adhere to the above timelines and ensure that the stability and resiliency of the network is our absolute first priority.

From all of the staff here at Servers Australia, we apologise for the disruption to our customers and the impact from this incident. We encourage you to reach out with any questions or feedback to your account manager who will be able to seek answers from the most appropriate person within Servers Australia.

We remain committed to being Australia's most trusted and innovative hosting provider.

Yours Sincerely,

Jared Hirst, CEO

RFO

Report Number	SAU-2017100501
Incident Type	Sydney Network Outage – Multiple Locations
Incident Start	2017-10-04 01:00:00
Incident End	2017-10-04 (end time varies for different customers)
Classification	For use by Servers Australia customers only.

Timeline of Events

1:00am: Servers Australia NOC receive alerts of multiple servers within our Sydney network experiencing high amounts of packet loss

1:00am: Servers Australia staff begin investigations of the alerts

1:10am: Servers Australia staff escalate to Network Engineering on-call. Priority changed to High Impact. Network Engineering begin investigations.

1:10am: Status notice created.

1:20am: Network Engineers locate a switch within the core network with High CPU load creating packet loss through a section of the SY1 core network.

1:25am: Network Engineers trace high CPU load to an issue with Spanning Tree on a legacy VLAN within the SY1 core network. VLAN disabled. Packet loss continues.

1:30am: Network Engineers locate secondary fault with a backbone routing port between 2 core network devices rapidly transitioning between an Up and Down state (Flapping). Traffic is shifted to a secondary redundant path by means of port shutdown. Diagnostic information shows faulty Fibre Optic Module.

1:35am: Network Engineers confirm stability of network operating on redundant path. Legacy VLAN remains disabled.

1:40am: Network Engineer departs for SY1 with replacement Fibre Optic modules and additional parts.

1:40am: Servers Australia NOC staff continue to monitor network and relay status to oncall engineer at regular intervals. Analysis of impacted customers shows customers at SY4 offline.

4:00am: Network Engineer arrives at SY1

4:05am: Network Engineer replaces faulty Fibre Optic module and cabling between devices.

4:10am: Network engineers shift traffic back to primary path and confirm hardware replacement resolved issue with the flapping port.

4:15am: Network Engineers implement workaround for the affected VLAN causing the spanning tree issues. VLAN enabled.

4:20am: Support staff confirm all services apart from Equinix SY4 online

4:30am: Network Engineers confirm issue is located at the SY4 end.

4:30am: Network Engineers unable to connect to Out Of Band management devices in SY1 or SY4 due to issue with ADSL connection used for Out Of Band.

4:32am: Network engineers leave Equinix SY1 to attend SY4

4:35am: Network Engineer arrives at SY4

4:40am: Network Engineers note the switching stack had reloaded and the stack had not returned to normal operation. Diagnostic information gathered for vendor.

4:50am: Network Engineers complete gathering of diagnostic information. Engineers find that no interface configuration existed on the switches due to the switches reversing the stack order upon reload of the switches.

5:20am: Network Engineers complete restoration of configuration from current backups on the SY4 core switches, restoring all connectivity to SY4 customers.

5:25am: NOC staff confirm normalised conditions. Monitor alerts clear for all services network wide. Priority downgrade to Normal Conditions.

5:45am: Required 20 minute cool down complete. Engineers depart site. All services normal. Status notice closed.

6:55am: NOC staff receive notifications of offline servers with Sydney datacentres and escalate back to on-call Network Engineer. Priority change to High Impact.

7:00am: Network Engineer begins investigations and notices that most services in Sydney offline. Priority change to Critical. Critical Actions invoked. Second network engineer, Operations Manager, Infrastructure Manager and CEO notified.

7:10am: All staff notified by NOC staff. Staff head to central location to begin situation management. Status notice created.

7:10am: NOC staff under guidance of Network Engineer oncall opens Smart Hands case with Equinix staff.

7:30am: Network Engineers arrive at Tuggerah office to continue investigation of issue

7:45am: Smart hands arrive onsite and communicate directly with Network Engineers.

8:00am: Issue located with a legacy Cisco core switch serving a small set of customers and completing some BGP routing.

8:15am: Cisco switch reloaded by smart hands as console unresponsive.

8:20am: Reload complete. No change to console. Technicians depart with spare switches and controllers. Smart hands directed to source secondary console device.

8:20am: Monitoring alerts appear for wider group of customers including Melbourne. Status notice updated.

8:30am: Smart hands source secondary console device. Console responsive on the switch.

8:35am: Cisco Switch confirmed to have reloaded with no configuration. Engineers restore minimal connectivity using an older configuration (external failsafe config) saved on the Cisco switch external flash.

8:45am: Cisco switch restores fail safe connectivity and basic connectivity confirmed. Outage alerts begin to clear.

8:55am: Engineers confirm only customers now affected by this outage are those directly connected to this Cisco switch

9:15am: Engineers confirm that Melbourne issue resolved.

9:20am: Engineers bring up production management network allowing access into the configuration backup repositories. Engineers start configuration merge between latest backup and failsafe config to restore the latest configuration without further downtime. 9:25am: Customers servers start to come back online as configuration merge takes place, engineers continue monitoring switch and wider network as restoration progresses

9:35am: Engineers observe high CPU usage on switch and rule the cause to be the configuration merge

9:55am: Configuration merge completes between failsafe configuration and the backup configuration. Switch confirmed to have full backup configuration. High CPU condition remains apparent. Investigations continue.

10:15am: Network Engineers locate a broadcast storm between the legacy Cisco core switch and the newer core network

10:25am: Engineers isolate the broadcast storm to a single VLAN. Cause found to be the temporary work around put in place at 4:15am.

10:30am: Engineers remove loop entirely and restored connectivity to the majority of services.
10:30am: NOC staff confirm normalised network conditions. Monitor alerts clear for the majority of services network wide. Priority downgrade to Normal Conditions.

10:35am: Network Engineers continue to work through a list of services still reporting issues.

3:00pm: Restored complete connectivity for all services.