HSWAW Power Incident (Incident/Postmortem #2)

Date: 2020/11/09
Status: Draft, **Review**, Public
Author: q3k@hackerspace.pl

Summary: power failure at Wolność 2a (W2A) caused unavailability of some services for up to 4 hours due to equipment/infrastructure cold boot issues.

Impact: SLI for colocation services affected. Internal services affected.

Services affected:

- boston-packets.hackerspace.pl, including: lists, email, ldap, IRC.
- k0.hswaw.net k8s cluster, including: matrix
- colocated equipment in dcr03 (two customers)

Root cause: Upstream power failure triggers all equipment on W2A to restart, multiple problems (see: what went wrong) caused a tail of service availability issues.

On call: q3k (pager), inf (noticed issues, helped with incident)

Lessons Learned

What went well

All of the bgp.wtf infrastructure (Internet customers, routers, distribution switches, CPEs) fully recovered immediately after the power outage.

What went wrong

B-class circuit breaker for Hackerspace circuit (not server room) tripped due to inrush current, disabling all of Hackerspace equipment (customs.hackerspace.pl, IOT, access points, door lock, etc). We do not have remote access to this breaker and had to resort to physical presence to fix it.

Misconfiguration of edge01 (nixos configuration not set to boot) caused BGP metrics exporting to fail, inducing a page; also caused DHCP server to fail preventing dcr01s{22,24} (k0 members) from acquiring a network address after startup.

Nodes in bc01 (eg. n01-03 which are k0 members) failed to start up due to being stuck in RAID controller error prompt (**“**press enter to continue booting”)

boston-packets.hackerspace.pl failed to automatically boot because of initramfs issues: not being able to find zroot/ROOT/gentoo and mount it.

dcr03sw01 (ToR switch in dcr03) failed to boot due to corrupted/broken flash, disconnecting all dcr03 colo. We did not have OOB access to the switch control plane and had to resort to physical presence to fix it.

Cooling air turbine for W2A colo did not start up after power failure, causing W2A server room temperature to hit 35 degrees. We do not have remote access to this turbine and had to resort to physical presence to fix it.

Access to KVM on bc01 nodes still sucks and took 15 minutes to get going.

Monitoring didn't page q3k to notify about issues downstream from edge01 - no monitoring of k0, boston-packets, customer devices, or anything else.

Where we got lucky

Multiple people volunteered to help, including being physically present at the space. If this wasn't the case we would've been kinda screwed, especially with cooling breaking down.

q3k didn't get woken up by inf's ping over Signal, but got woken up by a serendipitously spurious pagerduty alert.

inf was able to access the space through a *backdoor* even though the electronic lock died with the circuit breaker (UPS battery ran out after an hour or so?).

Action Items

Who	Type	What	Status
q3k	mitigate	Introduce monitoring and alerting for NixOS machines running configurations that are not set to boot.	accepted
enleth	prevent	Replace B-class circuit breaker on hackerspace circuit with a C/D-class breaker.	
q3k	prevent	Look into configuring the RAID controllers on M610 bc01 nodes to not stop boot on power failure, or work around this.	accepted
q3k	mitigate	Get OOB access to ToR switch in dcr03, possibly replacing ToR.	accepted
enleth	prevent	Build a remote control system for the cooling air turbine.	
q3k	mitigate	Build better documentation for KVM access to bc01, and/or fix software mess for iDRAC/KVM proxy.	accepted
implr	prevent	Fix initramfs on boston-packets.	
q3k	mitigate	Finish setting up monitoring for k0.hswaw.net.	accepted
q3k	mitigate	Set up monitoring for packet flow on dcr03 ToR.	accepted
q3k	mitigate	Set up alerting for boston-packets.hackerspace.pl.	accepted
implr	prevent	Ensure boston-packets can handle a reboot.	
q3k	other	Calculate impact on colocation SLO.	accepted
enleth	prevent	Check the door lock UPS to make sure it can handle an extended power failure.	

Timeline

2020/11/09, all times CET:

08:16:23 - last monitoring data from W2A scraped by monitoring.

Power outage . Incident begins.

08:19:46 - edge01.waw.bgp.wtf starts back up. All bgp.wtf services restored.

Power outage ends. Tail of affected services begins.

09:11:xx - inf pings q3k over signal (symptoms: edge01 up, but everything else dead)

09:16:xx - pagerduty fires for bgp prefix count alerts on edge01.waw, as metric scraping was still down

09:22:18 - q3k gets on staff IRC channel, debugging begins

09:40:54 - q3k discovers DHCP issue serving DHCP requests to dcr01s{22,24}

09:47:29 - q3k fixes configuration on edge01, DHCP and monitoring fixed, dcr01s{22,24} finish booting.

10:01:33 - inf arrives at W2A, notices electronic lock is unresponsive

10:06:11 - q3k discovers boston-packets is stuck on ZFS mount in initramfs.

10:08:41 - q3k discovers bc01n{01,02,03} stuck in RAID controller prompt, manually continues boot.

10:09:09 - inf enters hackerspace through *backdoor*, discovers and flips circuit breakers.

10:17:38 - q3k ensures k0 is up (etcd and crdb up)

10:24:40 - q3k begins looking into boston-packets ZFS issue

10:31:25 - q3k pages implr for boston-packets issues

10:32:51 - q3k manually imports ZFS pool on boston-packets and continues boot.

10:41:xx - q3k gets pages about temperature in W2A server room

10:48:xx - q3k gets in touch with W2A tenant to restore air cooling turbine.

10:58:56 - q3k gets notified about customer about colocation issues

11:06:21 - q3k confirms dcr03 ToR is likely dead

11:30:50 - inf arrives at space again and checks dcr03 ToR: supervisor engine light orange, linecard red

11:42:33 - inf plugs connects dcr03 ToR serial

11:46:14 - q3k manually boots stuck dcr03 switch

11:46:14 - customer confirms colo connectivity is back up

11:48:53 - q3k fixed dcr03 ToR boot issue (conf t; boot system flash bootflash:xxx.bin), restarts switch to check that it boots fine

11:50:19 - inf discovers PHP issues on boston-packets

11:52:49 - inf diagnoses and fixes PHP issues

11:55:47 - inf discovers owncloud issues

11:59:29 - inf diagnoses owncloud issues to be routing issues when reaching Ceph RadosGW

12:07:35 - q3k begins diagnosing owncloud routing issues

12:11:47 - q3k diagnoses owncloud routing issues to be a networking misconfiguration on boston-packets, fixes misconfiguration

Incident over.

Supporting information

Traffic dashboard on edge01.waw shows downtime between power incident and edge01 reconfiguration.