On September 10th, 2020 we experienced a series of events which caused significant disruption to our customer base. In this document we'll be outlining what happened, and what steps we will be taking to avoid such an incident in the future.

The Incident

At around 15:00 Pacific Time our NOC was alerted to a total outage in our Fremont datacenter. Investigation confirmed that our edge router was not pingable and our transit BGP sessions were offline. Using an out of band management connection, our team quickly determined the reason for the outage was a locked up hypervisor which was hosting our software-defined router. As the hypervisor was completely non-responsive, the call was made to forcibly reboot it through IPMI – at this point the true scale of the disaster became clear.

Not only did the hypervisor fail to boot, an error message complaining about a missing RAID array was being shown. The hypervisor was fully powered down, then powered back on, in the hopes that the missing array would be displayed; it was not and next steps needed to be taken.

We had the on site datacenter techs re-seat the drives, and eventually move the drives over to another server. We quickly determined that all but two of the drives were being detected in the second server, but the two drives making up the most crucial data – our customer database, website, automation server and clientarea code base could not be found. Making the situation worse, backups had been misconfigured and had been silently failing for 3 months. The last backup we had was from our old clientarea, before switching to our in house code base.

Immediate Remediation Steps

At this point we determined we would have to get as many customers online as possible, and deal with the failed array later. We quickly imported the functioning drives into other hypervisors and manually mapped the stored disk images to new VMs. We stood up a temporary router, using some development automation code that was stored on a different environment and placed a splash page on our website. At this point, all customers who were not on the failed array were back online. It was well past midnight local time and we called it a night.

The next day, we made arrangements to have our drives pulled and shipped back to us. The plan was to attempt data recovery on the array, and try to get as much data back as possible. We selected the fastest possible shipping option, however, customs and our courier would have other plans and it would be more then a week before we would actually see the drives.

In the meantime, we restored what we had – a clientarea 3 months out of date, a very beta version of our website and collection of emails, invoices and helpdesk tickets. With these scraps of information we began the painful process of piecing together a list of all our new customers and what services they had. We had also stood up a new ticking system, so our team would be able to track help requests more easily then via email. This is the state we would be in; manually searching and adding customers for the next week

Arrival of our Drives

On Monday, September 21 our drives would finally arrive after a number of delays. We verified the structural integrity of the connectors and tried to import the drives into a spare Dell server. We were unsuccessful at importing the drives, like with the servers at the datacenter, the drives were not being detected by the RAID controller.

Strangely, though the drives weren't detected by a Dell RAID controller, they could be detected on a Linux PC when connected to a standard SATA controller, however, because of the RAID formatting the filesystem could not be read. After a significant amount of research and trial we landed upon the open source application Testdisk – this marvelous piece of software was able to read our drives and restore all the saved virtual machine data.

With these revelations we were able to spend the next week uploading data over our measly 20mbps SOHO internet connection back to the Fremont datacenter. As virtual machines were uploaded, we notified customers their their machines were back online. We brought our in house billing system online and were able to import the new customer records into our legacy database and we were able to verify our code base was intact. Overall, things were looking promising once again.

The current State of Affairs

As we had already moved most of our customer base back to our legacy billing platform, we decided against moving immediately back to our in house system. We made the decision to import any erroneous records from our in house system back into our legacy system and make the legacy system our authoritative source of truth. This enabled two things: 1) it minimized the number of changes our customer base would have to endure in a short time, and 2) it allowed us to fix some long standing issues with our in house system without having any impact on our customers. The disadvantage, is that all the great service management options (such as VPS controls and IP leasing options) were made unavailable to our customers, but rest assured we are working on integrating these features into our legacy billing platform as soon as possible.

As things stand, services can be managed through several different URLs:

https://freerangecloud.com – our main website. Essentially a landing page with links to everything
https://legacy.freerangecloud.com – our billing system. All billing and service management will be conducted from this panel for the foreseeable future
https://clientarea.freerangecloud.com – Service control functions such as VPS control or IRR management
https://helpdesk.freerangecloud.com – Our ticketing system

As time goes on, we will begin integrating more services into our in house clientarea.

Take Away

This experience, while obviously less then ideal, has been a great learning experience for our small company. In particular we have identified a number of take-aways:

- **Ensure backups are being properly monitored**. If our backups had been monitored on a regular basis, this crisis would have been simple as pull the latest backup and restore it. We would have been up and running in less then 24 hours. We intend to modify our backup system to report every single job, not just failures. This means that a human will notice if the backups entirely stop running, since there will be no daily report. We also intend to implement regular testing of backups

- **Better communication with customers**. Throughout most of the crisis we were relying on social media, particularly Twitter to get information to our customers. This is less then ideal, and implementing a status page hosted outside our own infrastructure is a goal going forward.

- **Improved Monitoring**. Although it wouldn't have made much difference in this specific situation (since all our drives crapped at the same time), we intend on upping our monitoring game. In the event that one of our drives in our array had failed, we realized we likely wouldn't have been notified unless there was a serious performance degradation. We will be implementing strict monitoring and alerting of arrays based on SNMP data.

- **Separate development environment.** Our dev environment was hosted on the same infrastructure as our live environment, had it been hosted on a different hypervisor or in a different datacenter, even without backups we at least would have had a code base to fall back on. We intend on moving our dev environment to a totally separate datacenter, meaning even the most catastrophic failure would still give us a codebase to fall back on.