Post mortem on the Cloudflare Control Plane and Analytics Outage

11/04/2023

Matthew Prince

12 min read

This post is also available in 繁體中文, Français, Deutsch, Español, Português, 한국어, 简体中文 and 日本語.

Beginning on Thursday, November 2, 2023, at 11:43 UTC Cloudflare's control plane and analytics services experienced an outage. The control plane of Cloudflare consists primarily of the customer-facing interface for all of our services including our website and APIs. Our analytics services include logging and analytics reporting.

The incident lasted from November 2 at 11:44 UTC until November 4 at 04:25 UTC. We were able to restore most of our control plane at our disaster recovery facility as of November 2 at 17:57 UTC. Many customers would not have experienced issues with most of our products after the disaster recovery facility came online. However, other services took longer to restore and customers that used them may have seen issues until we fully resolved the incident. Our raw log services were unavailable for most customers for the duration of the incident.

Services have now been restored for all customers. Throughout the incident, Cloudflare's network and security services continued to work as expected. While

there were periods where customers were unable to make changes to those services, traffic through our network was not impacted.

This post outlines the events that caused this incident, the architecture we hac place to prevent issues like this, what failed, what worked and why, and the changes we're making based on what we've learned over the last 36 hours.

To start, this never should have happened. We believed that we had high availability systems in place that should have stopped an outage like this, even when one of our core data center providers failed catastrophically. And, while many systems did remain online as designed, some critical systems had non-obvious dependencies that made them unavailable. I am sorry and embarrassed for this incident and the pain that it caused our customers and our team.

Intended Design

Cloudflare's control plane and analytics systems run primarily on servers in three data centers around Hillsboro, Oregon. The three data centers are independent of one another, each have multiple utility power feeds, and each have multiple redundant and independent network connections.

The facilities were intentionally chosen to be at a distance apart that would minimize the chances that a natural disaster would cause all three to be impacted, while still close enough that they could all run active-active redundant data clusters. This means that they are continuously syncing data between the three facilities. By design, if any of the facilities goes offline then the remaining ones are able to continue to operate.

This is a system design that we began implementing four years ago. While most of our critical control plane systems had been migrated to the high availability cluster, some services, especially for some newer products, had not yet been added to the high availability cluster.

In addition, our logging systems were intentionally not part of the high availability cluster. The logic of that decision was that logging was already a distributed problem where logs were queued at the edge of our network and then sent back to the core in Oregon (or another regional facility for customers using regional services for logging). If our logging facility was offline then analytics logs would queue at the edge of our network until it came back online. We determined that analytics being delayed was acceptable.

Flexential Data Center Power Failure

The largest of the three facilities in Oregon is run by Flexential. We refer to this facility as "PDX-DC04". Cloudflare leases space in PDX-04 where we house our largest analytics cluster as well as more than a third of the machines for our high availability cluster. It is also the default location for services that have not yet been onboarded onto our high availability cluster. We are a relatively large customer of the facility, consuming approximately 10 percent of its total capacity.

On November 2 at 08:50 UTC Portland General Electric (PGE), the utility company that services PDX-04, had an unplanned maintenance event affecting one of their independent power feeds into the building. That event shut down one feed into PDX-04. The data center has multiple feeds with some level of independence that can power the facility. However, Flexential powered up their generators to effectively supplement the feed that was down.

Counter to best practices, Flexential did not inform Cloudflare that they had failed over to generator power. None of our observability tools were able to detect that the source of power had changed. Had they informed us, we would have stood up a team to monitor the facility closely and move control plane services that were dependent on that facility out while it was degraded.

It is also unusual that Flexential ran both the one remaining utility feed and the generators at the same time. It is not unusual for utilities to ask data centers to

drop off the grid when power demands are high and run exclusively on generators. Flexential operates 10 generators, inclusive of redundant units, capable of supporting the facility at full load. It would also have been possible for Flexential to run the facility only from the remaining utility feed. We haven't got a clear answer why they ran utility power and generator power.

Informed Speculation On What Happened Next

From this decision onward, we don't yet have clarity from Flexential on the root cause or some of the decisions they made or the events. We will update this post as we get more information from Flexential, as well as PGE, on what happened. Some of what follows is informed speculation based on the most likely series of events as well as what individual Flexential employees have shared with us unofficially.

One possible reason they may have left the utility line running is because Flexential was part of a program with PGE called DSG. DSG allows the local utility to run a data center's generators to help supply additional power to the grid. In exchange, the power company helps maintain the generators and supplies fuel. We have been unable to locate any record of Flexential informing us about the DSG program. We've asked if DSG was active at the time and have not received an answer. We do not know if it contributed to the decisions that Flexential made, but it could explain why the utility line continued to remain online after the generators were started.

At approximately 11:40 UTC, there was a ground fault on a PGE transformer at PDX-04. We believe, but have not been able to get confirmation from Flexential or PGE, that this was the transformer that stepped down power from the grid for the second feed that was still running as it entered the data center. It seems likely, though we have not been able to confirm with Flexential or PGE, that the ground

fault was caused by the unplanned maintenance PGE was performing that impacted the first feed. Or it was a very unlucky coincidence.

Ground faults with high voltage (12,470 volt) power lines are very bad. Electric systems are designed to quickly shut down to prevent damage when one occurs. Unfortunately, in this case, the protective measure also shut down all of PDX-04's generators. This meant that the two sources of power generation for the facility — both the redundant utility lines as well as the 10 generators — were offline.

Fortunately, in addition to the generators, PDX-04 also contains a bank of UPS batteries. These batteries are supposedly sufficient to power the facility for approximately 10 minutes. That time is meant to be enough to bridge the gap between the power going out and the generators automatically starting up. If Flexential could get the generators or a utility feed restored within 10 minutes then there would be no interruption. In reality, the batteries started to fail after only 4 minutes based on what we observed from our own equipment failing. And it took Flexential far longer than 10 minutes to get the generators restored.

Attempting to Restore Power

While we haven't gotten official confirmation, we have been told by employees that three things hampered getting the generators back online. First, they needed to be physically accessed and manually restarted because of the way the ground fault had tripped circuits. Second, Flexential's access control system was not powered by the battery backups, so it was offline. And third, the overnight staffing at the site did not include an experienced operations or electrical expert — the overnight shift consisted of security and an unaccompanied technician who had only been on the job for a week.

Between 11:44 and 12:01 UTC, with the generators not fully restarted, the UPS batteries ran out of power and all customers of the data center lost power. Throughout this, Flexential never informed Cloudflare that there was any issue at

the facility. We were first notified of issues in the data center when the two routers that connect the facility to the rest of the world went offline at 11:44 UTC. When we weren't able to reach the routers directly or through out-of-band management, we attempted to contact Flexential and dispatched our local tea. to physically travel to the facility. The first message to us from Flexential that they were experiencing an issue was at 12:28 UTC.

> *We are currently experiencing an issue with power at our [PDX-04] that began at approximately 0500AM PT [12:00 UTC]. Engineers are actively working to resolve the issue and restore service. We will communicate progress every 30 minutes or as more information becomes available as to the estimated time to restore. Thank you for your patience and understanding.*

Designing for Data Center Level Failure

While the PDX-04's design was certified Tier III before construction and is expected to provide high availability SLAs, we planned for the possibility that it could go offline. Even well-run facilities can have bad days. And we planned for that. What we expected would happen in that case is that our analytics would be offline, logs would be queued at the edge and delayed, and certain lower priority services that were not integrated into our high availability cluster would go offline temporarily until they could be restored at another facility.

The other two data centers running in the area would take over responsibility for the high availability cluster and keep critical services online. Generally that worked as planned. Unfortunately, we discovered that a subset of services that were supposed to be on the high availability cluster had dependencies on services exclusively running in PDX-04.

In particular, two critical services that process logs and power our analytics — Kafka and ClickHouse — were only available in PDX-04 but had services that

depended on them that were running in the high availability cluster. Those dependencies shouldn't have been so tight, should have failed more gracefully, and we should have caught them.

We had performed testing of our high availability cluster by taking each (and both) of the other two data center facilities entirely offline. And we had also tested taking the high availability portion of PDX-04 offline. However, we had never tested fully taking the entire PDX-04 facility offline. As a result, we had missed the importance of some of these dependencies on our data plane.

We were also far too lax about requiring new products and their associated databases to integrate with the high availability cluster. Cloudflare allows multiple teams to innovate quickly. As such, products often take different paths toward their initial alpha. While, over time, our practice is to migrate the backend for these services to our best practices, we did not formally require that before products were declared generally available (GA). That was a mistake as it meant that the redundancy protections we had in place worked inconsistently depending on the product.

Moreover, far too many of our services depend on the availability of our core facilities. While this is the way a lot of software services are created, it does not play to Cloudflare's strength. We are good at distributed systems. Throughout this incident, our global network continued to perform as expected. While some of our products and features are configurable and serviceable through the edge of our network without needing the core, far too many today fail if the core is unavailable. We need to use the distributed systems products that we make available to all our customers for all our services, so they continue to function mostly as normal even if our core facilities are disrupted.

Disaster Recovery

At 12:48 UTC, Flexential was able to get the generators restarted. Power returned to portions of the facility. In order to not overwhelm the system, when power is restored to a data center it is typically done gradually by powering back on one circuit at a time. Like the circuit breakers in a residential home, each customer serviced by redundant breakers. When Flexential attempted to power back up Cloudflare's circuits, the circuit breakers were discovered to be faulty. We don't know if the breakers failed due to the ground fault or some other surge as a result of the incident, or if they'd been bad before, and it was only discovered after they had been powered off.

Flexential began the process of replacing the failed breakers. That required them to source new breakers because more were bad than they had on hand in the facility. Because more services were offline than we expected, and because Flexential could not give us a time for restoration of our services, we made the call at 13:40 UTC to fail over to Cloudflare's disaster recovery sites located in Europe. Thankfully, we only needed to fail over a small percentage of Cloudflare's overall control plane. Most of our services continued to run across our high availability systems across the two active core data centers.

We turned up the first services on the disaster recovery site at 13:43 UTC. Cloudflare's disaster recovery sites provide critical control plane services in the event of a disaster. While the disaster recovery site does not support some of our log processing services, it is designed to support the other portions of our control plane.

When services were turned up there, we experienced a thundering herd problem where the API calls that had been failing overwhelmed our services. We implemented rate limits to get the request volume under control. During this period, customers of most products would have seen intermittent errors when making modifications through our dashboard or API. By 17:57 UTC, the services that had been successfully moved to the disaster recovery site were stable and

most customers were no longer directly impacted. However, some systems still required manual configuration (e.g., Magic WAN) and some other services, largely related to log processing and some bespoke APIs, remained unavailable until we were able to restore PDX-04.

Some Products and Features Delayed Restart

A handful of products did not properly get stood up on our disaster recovery sites. These tended to be newer products where we had not fully implemented and tested a disaster recovery procedure. These included our Stream service for uploading new videos and some other services. Our team worked two simultaneous tracks to get these services restored: 1) reimplementing them on our disaster recovery sites; and 2) migrating them to our high-availability cluster.

Flexential replaced our failed circuit breakers, restored both utility feeds, and confirmed clean power at 22:48 UTC. Our team was all-hands-on-deck and had worked all day on the emergency, so I made the call that most of us should get some rest and start the move back to PDX-04 in the morning. That decision delayed our full recovery, but I believe made it less likely that we'd compound this situation with additional mistakes.

Beginning first thing on November 3, our team began restoring service in PDX-04. That began with physically booting our network gear then powering up thousands of servers and restoring their services. The state of our services in the data center was unknown as we believed multiple power cycles were likely to have occurred during the incident. Our only safe process to recover was to follow a complete bootstrap of the entire facility.

This involved a manual process of bringing our configuration management servers online to begin the restoration of the facility. Rebuilding these took 3 hours. From there, our team was able to bootstrap the rebuild of the rest of the

servers that power our services. Each server took between 10 minutes and 2 hours to rebuild. While we were able to run this in parallel across multiple servers, there were inherent dependencies between services that required some to be brought back online in sequence.

Services are now fully restored as of November 4, 2023, at 04:25 UTC. For most customers, because we also store analytics in our European core data centers, you should see no data loss in most analytics across our dashboard and APIs. However, some datasets which are not replicated in the EU will have persistent gaps. For customers that use our log push feature, your logs will not have been processed for the majority of the event, so anything you did not receive will not be recovered.

Lessons and Remediation

We have a number of questions that we need answered from Flexential. But we also must expect that entire data centers may fail. Google has a process where when there's a significant event or crisis they can call a Code Yellow or Code Red. In these cases, most or all engineering resources are shifted to addressing the issue at hand.

We have not had such a process in the past, but it's clear today we need to implement a version of it ourselves: Code Orange. We are shifting all non-critical engineering functions to focusing on ensuring high reliability of our control plane. As part of that, we expect the following changes:

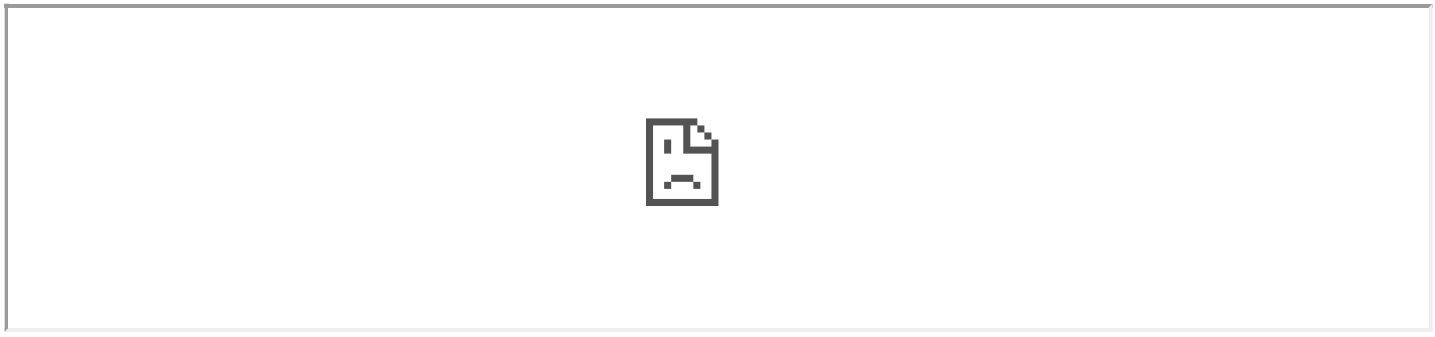
- Remove dependencies on our core data centers for control plane configuration of all services and move them wherever possible to be powered first by our distributed network
- Ensure that the control plane running on the network continues to function even if all our core data centers are offline

- Require that all products and features that are designated Generally Available must rely on the high availability cluster (if they rely on any of our core data centers), without having any software dependencies on specific facilities
- Require all products and features that are designated Generally Available have a reliable disaster recovery plan that is tested
- Test the blast radius of system failures and minimize the number of services that are impacted by a failure
- Implement more rigorous chaos testing of all data center functions including the full removal of each of our core data center facilities
- Thorough auditing of all core data centers and a plan to reaudit to ensure they comply with our standards
- Logging and analytics disaster recovery plan that ensures no logs are dropped even in the case of a failure of all our core facilities

As I said earlier, I am sorry and embarrassed for this incident and the pain that it caused our customers and our team. We have the right systems and procedures in place to be able to withstand even the cascading string of failures we saw at our data center provider, but we need to be more rigorous about enforcing that they are followed and tested for unknown dependencies. This will have my full attention and the attention of a large portion of our team through the balance of the year. And the pain from the last couple of days will make us better.

To learn more about our mission to help build a better Internet, start here. If you're looking for a new career direction, check out our open positions.

Discuss on Hacker News



Outage Post Mortem

Follow on X

Matthew Prince | @eastdakota
Cloudflare | @cloudflare

RELATED POSTS

November 01, 2023 11:39 AM

Cloudflare incident on October 30, 2023

Multiple Cloudflare services were unavailable for 37 minutes on October 30, 2023, due to the misconfiguration of a deployment tool used by Workers KV....

By Matt Silverlock, Kris Evans

Post Mortem

October 20, 2023 4:39 PM

How Cloudflare mitigated yet another Okta compromise

On Wednesday, October 18, 2023, we discovered attacks on our system that we were able to trace back to Okta. We have verified that no Cloudflare customer information or systems were impacted by this event because of our rapid response. ...

By Sourov Zaman, Lucas Ferreira, Kimberly Hall, Grant Bourzikas

Okta, Post Mortem, 1.1.1.1

October 04, 2023 2:40 PM

1.1.1.1 lookup failures on October 4, 2023

On 4 October 2023, Cloudflare experienced DNS resolution problems. Some users may have received SERVFAIL DNS responses to valid queries. In this blog, we're going to talk about what the failure was, why it occurred, and what we're doing to make sure this doesn't happen again...

By Ólafur Guðmundsson

1.1.1.1, Post Mortem, Outage

August 02, 2023 8:05 AM

Hardening Workers KV

A deep dive into the recent incidents relating to Workers KV, and how we're going to fix them...

By Matt Silverlock, Charles Burnett, Rob Sutter, Kris Evans

Cloudflare Workers, Reliability, Post Mortem