Post-Incident Analysis

Broadband Network Outage on 10/24/2022

Summary

At about 7:00pm on Monday, October 24th, 2022, there was an unexpected loss of internet service for 63 confirmed customers – but possibly more. Service was restored at 4:35pm on Tuesday, October 25, 2022. This outage was inconsistent with the level of service we strive to provide, and it is imperative we identify and communicate the root cause to prevent similar events in the future. What follows is an analysis of the event, what was done to fix it, and what we are doing to make meaningful improvements to prevent it from happening again.

Overview of Concord Broadband

Concord Broadband is an enterprise in the Town of Concord that falls under the umbrella of the Concord Municipal Light Plant (CMLP), the Town's electric utility. Servicing around 1,600 residential and commercial customers, Concord Broadband also provides internet and network connectivity to all of Concord's municipal buildings through its fiber infrastructure and a combination of three separate internet service providers (ISPs). At the time of this report, Concord Broadband consists of two full-time Telecom Technicians, a Senior Telecom Technician, and two full-time Network Engineers (one currently vacant) who report to the Broadband Manager. The Manager reports to the Chief Technology Officer (CTO), who is overseen by the Town Manager.

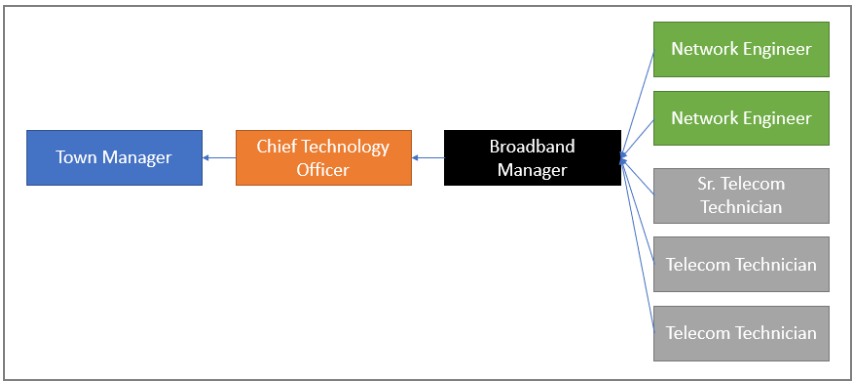

Figure 1: Organizational Chart for Concord Broadband

A third-party call center (Netegrity) handles basic troubleshooting service for customers over the phone. The equipment, including ONTs deployed in businesses and residences, are manufactured and supported by the vendor, Calix and covered by a 24/7 service window. The networking equipment in use, manufactured by Cisco, is also covered by 24x7 service.

More information about the department, its offerings and current speeds and pricing can be found on its web page: https://www.concordma.gov/broadband

Incident Background

In December of 2021, rodents chewed through several fibers (7, ranging from 12-strand to 96-strand) adjacent to the Laws Brook LCC[1]. Emergency repairs were conducted at the time to get impacted customers online as soon as possible. Due to the nature of the damage to the cables, it was clear that the cables would have to be removed, replaced, and re-spliced before the winter of 2022 to ensure they would withstand the elements in the long-term.

Arrangements were made to have a third-party assist with this work, and after several delays by the vendor, it took place September 27-30 and October 19-20, 2022.

To perform these repairs, a cable is cut, immediately suspending service for all customers downstream, and then the cable is carefully removed. A new cable is put in its place, and that cable's fiber is spliced at the respective splice case and terminated at the LCC.

The first four damaged cables were removed with only an impact of the customers utilizing those fibers, ranging from under 10 to just over 20 with a total of 55. But on the final night of splicing, the cable cut was the feeder fiber, and something different happened.

The feeder fiber at the LCC carries all the traffic to the customers that branch out at that location. When it is cut, *all* the traffic is suspended to the port in a piece of networking equipment in our data center configured to provide that service. This event is different from a handful of a customers, and it trips certain logic in the equipment to ensure the health of the network.

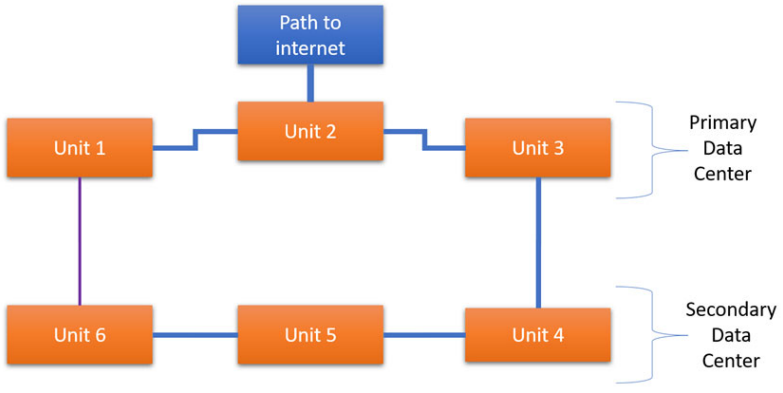

Figure 2: Ring architecture of Broadband equipment

[1] LCC: Local convergence cabinet, where cables are brought in to patch panels to be used to connect customers back to head-end equipment in data centers.

Pictured above, Units 1-6 are the devices that provide internet connectivity to all customers. They have redundant links, and the connection between 1 and 6 (in purple) automatically becomes active if the link between Unit 4 and Unit 3 goes down. They operate in a carefully choreographed ring, constantly reporting information to one another to monitor for an issue that would require them to failover into a new configuration to keep all possible customers connected to the internet.

The cable replacement work was performed Thursday night, and customers were back up at 1:30am. Later that morning, though, we had a handful of customers report they were offline with no internet connectivity. These customers all had a configuration in common (and all were on Unit 4 in the diagram above), so we made a change to that configuration, and the customers came back up.

Knowing that this resulted without any employee interaction, we called Calix, the vendor that makes these Units (officially known as “Modular Systems”), to have them look into this issue.

Because internet service provider networking is complex, we have three different vendors responsible for equipment that provides Broadband service for customers, and when there is a general networking issue, all have to be involved. Since at that time all of our customers were back up and running, the call back time for each provider was between 1-4 hours. We worked with them to create a testing regimen to isolate each solution. This work continued over the weekend, with our Network Engineer and Telecom Technician working until midnight on Friday, after 10pm on Saturday, and 7pm on Sunday. We prioritized this work because even though we had a handle on the affected customers, we knew there was an unexplained underlying network issue.

We prioritized this work because even though we had a handle on the affected customers, we knew there was an unexplained underlying network issue.

On Monday morning we reengaged Calix with findings that every other piece of equipment had been tested and was not found to be involved in this issue, and they continued their support escalations to the highest level. Developers of their software were involved as well to see if some part of the code that runs on cards or shelfs (components of each Unit) could be to blame for this.

At the end of the day, the official suggestion from Calix was to try moving a port (carrying 20-50 customers) to another card on a different Unit to see if the issue went away. Before we did that, however, they wanted to consult their team of professionals to vet this recommendation, and they needed Concord Broadband to choose which port would be most ideal based on the lowest number of customers impacted. Their ultimate thought was that one of the two cards on Unit 4 was the culprit and needed to be reseated or replaced. This work would be scheduled, most likely, for a maintenance window on Tuesday night.

We began doing that preparation work remotely until 7:00pm on Monday when we received a call from our helpdesk provider saying that there was a "widespread outage."

Incident

Nothing precipitated the outage. No configuration change was made, and no employees were logged into the system. Nothing was rebooted, and no actions by vendors or employees led to the sudden outage.

There are two types of outages in the Fiber-to-the-Home world: Those that involve physical issues (with fiber being broken/cut or customer equipment breaking/losing power) and those that involve networking issues. When the former occurs, our staff is instantly notified, and we have immediate visibility into the number of impacted customers and the degree of the outage. But when networking issues occur, notifications can be difficult.

For example, while staff are alerted if traffic dips below a certain threshold or links go down, these alerts aren't tripped if outages are small or inconsistent. In this case, a sudden influx of calls to the helpdesk triggered the "widespread outage" response, and two personnel immediately left for our primary data center.

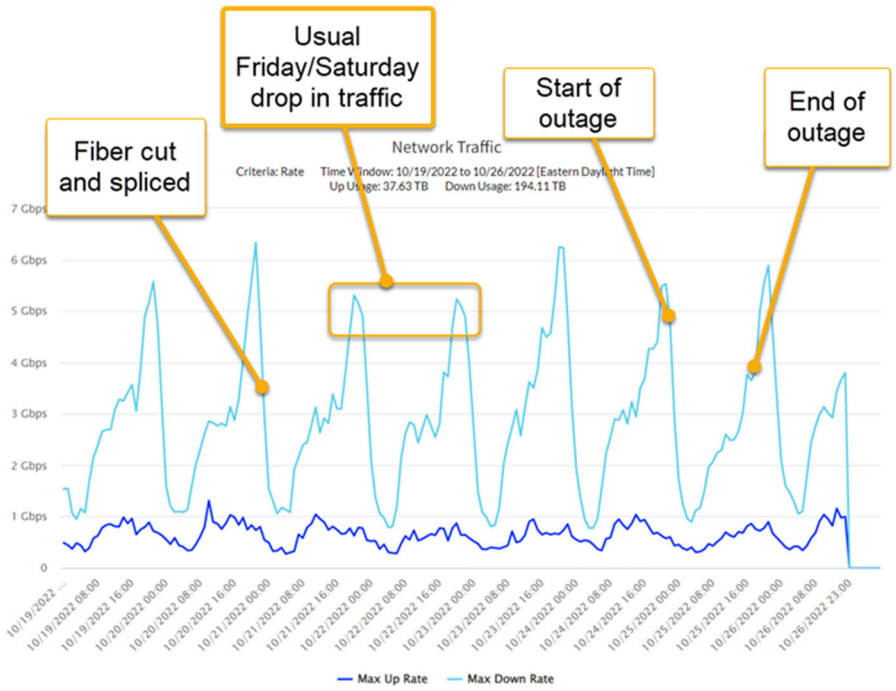

Figure 3: Network activity chart from 10/19/2022 - 10/26/2022

While on-site, it was clear that, while we had 25 calls, the outage was not technically widespread. We learned this from the level of overall traffic we were seeing on the network, and other metrics we can gather. We began by testing all other equipment again to ensure that no router, switch or DHCP appliance was the issue.

Our helpdesk vendor has between 6-10 technicians on call at any given time. Because large outages can quickly overwhelm their call center, they typically put a recorded message in place alerting customers that there is a known issue under investigation. This is a common practice across the call center helpdesk industry because they cannot reasonably staff to levels necessary to handle peak call volumes during large outages. This case was no different, and by around 7:30pm, Netegrity enabled a recorded message telling users that there was a known outage under investigation.

With customers clearly impacted, we decided to reboot the card on Unit 4, which had been Calix's planned remediation step if moving the port didn't work. We had a handful of customer names who were impacted at 7pm, and we also had an ONT we were using to test in the field. After the reboot, the test ONT was back online, and the customers we called back were online as well.

Systems looked normal, so we reached out to our helpdesk vendor to ask they remove the recorded message to ensure that any impacted customers could get through to an operator. Concord Broadband employees packed up, and after 30 minutes went home around midnight.

Within 30 more minutes, it was clear that people were still having trouble, and that the card reboot had not solved all of the problems. We put out additional notices on Twitter, directing people to email us if they were having issues, and we promptly engaged Calix to begin triaging.

Our case was constantly escalated to subsequent levels after several rounds of initial troubleshooting. While this was happening, customers began emailing, and this provided data that could be used to identify a discernable pattern as to who was impacted. It was clear after several emails that it was customers on all 6 Units in the diagram, some with static IP addresses and some with dynamic IP addresses (though a good deal higher percentage of static customers).

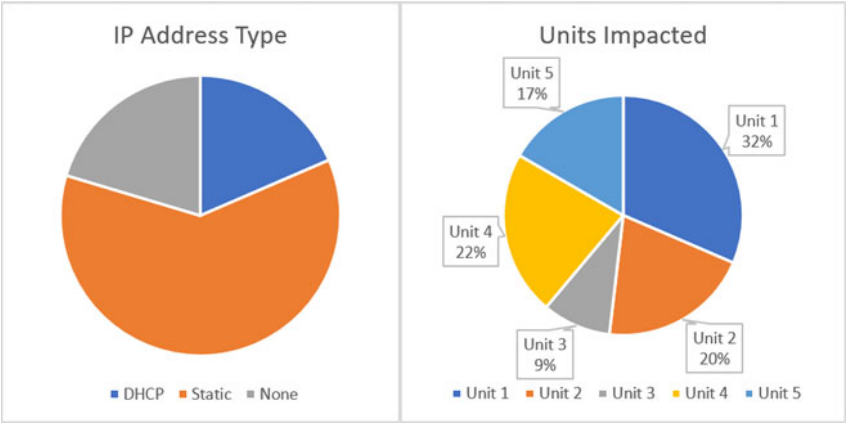

Figure 4: Data from customers enabled us to identify patterns in real time.

An employee from Concord Broadband remained on the phone all night long with our vendor to continue troubleshoot the issue to try to bring about a resolution. While there were talented support personnel from Calix available all night long, they do not have as many resources at the highest level of escalation

overnight, so there were few people willing to create an action plan. They continued testing and tried to change the flow of traffic around the ring. Occasionally this works to resolve connectivity issues like we were seeing. While we always reminded them that this event began after a port on a card went down due to a fiber repair, they always dismissed that as an unlikely root cause because the Units are designed to handle events like that. Many troubleshooting steps were taken, and the feedback from our customers was proving indispensable to confirm that their outage continued.

In the early morning, the case got to a team familiar with the issue from day one, and they had us troubleshoot and test the ring by doing packet captures from impacted ONTs to see where the data stopped flowing. Unfortunately, sometimes software says a piece of hardware is doing something when it is in fact not doing that thing, so these packet captures did not help identify *what* the problem was, but they did steer us toward *where* it was: Unit 2. As you can see from the charts above, 20% of impacted customers were on Unit 2, and since Unit 2 has a direction connection to the internet, the viability of the ring did not come into play, leading Calix to suspect there was an issue with that Unit.

During this time, CMLP's customer service representatives fielded dozens of phone calls from customers looking for updates. Concord Broadband's Telecom Technicians assisted with information gathering, correlation of data, and field testing.

Calix's team of engineers and developers continued to look into the logs and packet captures to determine if they could see any sign that Unit 2 was the source of the issue. The fear was that if Unit 2 was *not* causing this problem, then rebooting it may not only *not* remediate the issue, but it could cause a much larger outage, potentially impacting *all* customers (since Unit 2 is the source of the internet for all ONTs).

Throughout the night and day, Concord Broadband issued approximately 10 Tweets and posted 3 News and Notices announcements. We sent several emails to Town staff, management, our helpdesk vendor, and all the customers who had written or called in and provided an email addresses.

At around 4pm, after being on the phone and a screen share since midnight the day before, an action plan was developed to reboot Unit 2. The engineers from Calix asked what maintenance window we wanted to use, and we responded that we wanted to do it as soon as possible since we had customers – at this point 63 confirmed – down for almost 20 hours. Our other concern in waiting for a middle-of-the-night maintenance window was that it would leave fewer engineers and support personnel available from Calix if the situation deteriorated.

The earliest they would be available was at 4:30pm because they needed to run the action plan by their supervisors and colleagues and mobilize resources in the event the issue deteriorated. At 4:32pm, the command was issued to reboot Unit 2, and we watched dashboards showing customer traffic to understand the impact.

Within 2-4 minutes of issuing the command, we received an email from a customer saying she was back online. Other customers soon followed. We spot-checked customers from our confirmed list of impacted, and they all had IP addresses, and they were all passing bidirectional traffic.

We continued by returning some phone calls and testing every customer who had contacted us.

Analysis

We next moved into the analysis phase to try to understand why this particular Unit 2 had become corrupted and needed to be rebooted. Unlike your home PC that has a relatively low reliability when it comes to crashing, these units are designed to not need to be rebooted… ever. Every error is designed to be handled, and every software exception is designed to be self-correcting. It is not uncommon for the devices to only be rebooted when new firmware is installed, and the engineers we spoke with told us 4 years was not an uncommon up-time statistic.

With the Calix team, we explained again the work performed the week before and asked if it could have been related to this event or if was a coincidence. The assessment they gave was inconclusive.

Events like that, when a port goes down, should not under any circumstance impact the integrity of the ring or any single Unit. The fiber spliced, after all, was on Unit 5, and the first impacted customers were on Unit 4, yet the problem turned out being Unit 2.

It Concord Broadband's internal assessment that the splicing work on Thursday most likely led to this customer-impacting event. We have gathered all system logs from Unit 2 and passed them along to Calix for further analysis. Should they provide a more definitive root cause, we will update this document and inform our customers.

Past Remediations

Concord Broadband experienced a similar but much worse event in June of 2021. At the time, Town Leadership, including the Town Manager, Light Plant Director, CTO and Broadband Manager (then called the Telecom Director) issued a series of recommendations for improvements to both prevent future outages and offer quicker resolutions for those that did occur.

The primary recommendations were:

- Conduct additional training for Broadband personnel
- Put in place a more accurate and up-to-date flow analysis tool to allow staff to determine an estimated impact on customers when there is no other hard data.
- Remove the reliance on the Town's datacenter for DHCP (both because it could fail and because it complicated troubleshooting)

- Ensure all software is adequately licensed and all hardware is not end-of-life and carries active support contracts
- Update all network diagrams, which are needed at each round of triaging as a case is escalated
- Adopt an incident response plan that directs recovery efforts
- Provide regular, meaningful updates to internal stakeholders, the helpdesk vendor, and customers during outages.
- Ensure all hardware is up to date with its firmware/software, since troubleshooting often takes longer on older versions.
- Adopt the newest technology tools that would enable instant communication and collaboration across the remediation team and management.

We are happy to report that every single action item on the list was addressed ahead of this recent issue, and it played a huge role in ensuring that this outage did not extend for days. Some of those items kept staff working during maintenance windows during the night and felt unimportant in the moment, but everyone knew that the steps were needed to provide consistent 24x7 service to our customers.

Issues

What follows is a summary of the issues encountered during this incident. They are broken down into resource, technical, communication, and planning issues.

1. Resource
    a. Concord Broadband has positions for 2 Network Engineers. One of our Network Engineers left the organization as of September 1, and that position has not yet been filled. Having one more person available would have potentially allowed working staff a chance to rest for a few minutes or provided additional troubleshooting support. (I will mention parenthetically that the Network Engineer who recently left *did* provide a few hours of remote support during this outage, which helped at times with our troubleshooting.)
    b. The vendor support at times, mainly from midnight through 4am was not on-par with what we would expect from them during an outage. While there was no down-time, there were also no meaningful recommendations from that team.
2. Technical
    a. Every piece of equipment was up-to-date and covered by support/maintenance. There was one part needed for troubleshooting that we had to collect from another site instead of having it at the datacenter (we had one of these for two data centers). While it didn't slow down troubleshooting since there were other steps we could take at the time, it would have been less stressful if we hadn't had to drive across town to retrieve that part. Another should be ordered.
3. Communication
    a. This is without doubt the biggest failure in the list. It is understandable that the helpdesk vendor cannot answer over one thousand calls simultaneously in the event of a total outage and may need to resort to a recorded message. But the message posted should contain

meaningful information and updates about the scope of the event, service restoration and where people can go for more information. Even though this information was posted on the website, not all residents and businesses knew to look there, so they called the number on the sticker on their ONTs or on their bills.

4. Planning
    a. While Concord Broadband developed an incident response plan in 2021, it should be regularly updated, and employees should receive routine training on it. Periodic table-top exercises should take place.

Action Items

The issues identified above are not difficult to remediate. As for a timeline, we anticipate:

Action Item	Time to resolution	People assigned
Hire 2nd Network Engineer	1-3 months	Broadband Manager, CTO
Improve vendor support	1-3 weeks	Broadband Manager
Purchase troubleshooting part	1 week	Network Engineer
Improve helpdesk communication	1-2 weeks	Broadband Manager
Begin training on Incident Response Plan and tabletop exercises	1 month	Broadband team

Next Steps

The next steps are to communicate what this post-incident analysis contains to all stakeholders. Concord Broadband will publish this update on the Town's website, make it available for the Select Board packet and discuss it at November's Light Board meeting.

Apology

Nobody likes to lose a vital service like their internet, and it is incredibly impactful on people's lives and on businesses when it happens. We sincerely apologize for this outage and its impact on you, our customers.

Response

Concord Broadband has been and always will be a community resource. What makes us different from competitors is not the services or speeds we offer, but the fact that we live and work in the community and

genuinely care about our customers. We try to let that inspire us every day so we can provide amazing service to all of you.

Throughout this outage there have been several staff members who have gone way above and beyond what any job should ask of an employee. We would like to thank Thomas Boadu for his non-stop contributions from the beginning of the issue on Friday morning, along with Todd Bagdasarian, Rob Muir, and Marc Goulet for their amazing help during this outage.

We would also like to thank the customers who were so incredibly supportive, with some writing:

- "Thank you for your follow up call."
- "Thanks for the hard work and diligence to get the service back up & running!"
- "Thank you for your continued updates yesterday."
- "Just want to thank you and your team for excellent work in resolving the issue and excellent customer care."
- "Thanks […] for great service and resolution of this issue"
- "Thanks so much!"
- "I appreciate all the communication."
- "We've got our service back! Thank you so much, it must have been a tough 24 hours."
- "I know you are all working very hard to fix the problem!"

Conclusion

It is our sincere goal to maintain a transparent path forward to rebuild trust with our customers after this incident. We would like to reiterate our apology for this incident and reaffirm our commitment to prioritize broadband as the critical infrastructure it is.

Should people like to discuss this document in greater detail, feel free to contact us at broadband@concordma.gov.