October 2021
Geoff Huston

Learning from Facebook's Mistakes

It was only a few weeks back, in July of this year, where I remarked that an Akamai report of an outage was unusual for this industry (https://www.potaroo.net/ispcol/2021-07/outage.html). It was unusual in that it was informative in detailing their understanding of the root cause of the problem, describing the response that they performed to rectify the immediate problem, the measures being undertaken to prevent a recurrence of this issue and the longer-term measures to improve the monitoring and alerting processes used within their platform.

At the time I noted that it would be a positive step forward for this industry if Akamai's outage report was not unusual in any way. It would be good if all service providers spent the time and effort post rectification of an operational problem to produce such outage reports as a matter of standard operating procedure. It's not about apportioning blame or admitting liability. It's all about positioning these services as the essential foundation our of digital environment and stressing the benefit of adopting a common culture of open disclosure and constant improvement as a way of improving the robustness of all these services. It's about appreciating that these days these services are very much within the sphere of public safety and their operation should be managed in the same way. We should all be in a position to improve the robustness of these services by appreciating how vulnerabilities can lead to cascading failures.

On October 4th Facebook managed to achieve one of the more impactful of outages of the entire history of the Internet, assuming that the metric of "impact" is how many users one can annoy with a single outage. In Facebook's case the 6-hour outage affected the services it provides so some 3 billion users, if we can believe Facebook's marketing hype.

So, what did we learn about this outage? What was the root cause? What were the short-term mitigations that they out in place? Why did it take more than 6 hours to restore service? (Yes, for a configuration change that presumably had a black-out plan that's an impressively long time!) What are they doing now to ensure that this situation won't recur? What can we as an industry learn from this outage to ensure that we can avoid a recurrence of such a widespread outage in other important and popular service platforms?

These are all good questions, and if we are looking for answers then Facebook's outage report is not exactly a stellar contribution. It's short enough for me to reproduce in its entirety here:

> To all the people and businesses around the world who depend on us, we are sorry for the inconvenience caused by today's outage across our platforms. We've been working as hard as we can to restore access, and our systems are now back up and running. The underlying cause of this outage also impacted many of the internal tools and systems we use in our day-to-day operations, complicating our attempts to quickly diagnose and resolve the problem.
>
> Our engineering teams have learned that configuration changes on the backbone routers that coordinate network traffic between our data centers caused issues that interrupted this communication. This disruption to network traffic had a cascading effect on the way our data centers communicate, bringing our services to a halt.
>
> Our services are now back online and we're actively working to fully return them to regular operations. We want to make clear that there was no malicious activity behind this outage — its root cause was a faulty configuration change on our end. We also have no evidence that user data was compromised as a result of this downtime. *(Updated on Oct. 5, 2021, to reflect the latest information)*

> People and businesses around the world rely on us every day to stay connected. We understand the impact that outages like these have on people's lives, as well as our responsibility to keep people informed about disruptions to our services. We apologize to all those affected, and we're working to understand more about what happened today so we can continue to make our infrastructure more resilient.
> https://engineering.fb.com/2021/10/04/networking-traffic/outage/

Yes, they are "sorry". Well, they could hardly say anything else could they.

Yes, they did this to themselves. Again, nothing unusual here, in that configuration changes are the most common cause of service faults. That's why most communications service providers impose a configuration freeze over the important periods, such as "Black Friday" in the US, or the new year holiday period, and that's why such freeze periods are typically the most stable of the entire year! But in Facebook's case whatever pre-installation tests they performed, if indeed they did any at all, failed to identify a risk in the change process. I guess the engineering team were still applying Mark Zuckerberg's operational mantra of moving fast and breaking things, and doing so with a little too much zeal.

And "they are working to understand more about what happened today so we can continue to make our infrastructure more resilient." No details.

I must admit this this report is a state-of-the-art example of a vacuous statement that takes four paragraphs to be largely uninformative.

NBC News reported that: "A Facebook employee said it appeared to be a problem with the Domain Name System, the "phone book" of the internet, which computers use to look up individual websites. "I wish I knew. No internal tooling, DNS seems totally borked. Everyone is just sort of standing around," the source said. "No reason at this point to suspect anything malicious, but the outage is affecting pretty much everything. Can't even access third-party tools." (https://www.nbcnews.com/tech/social-media/social-media-giants-facebook-instagram-whatsapp-all-experiencing-outages-n1280726).

It seems sad that this NBC report was far more informative than the corporate blather that Facebook posted as their statement from engineering.

What really did happen and what can we learn from this outage?

For this I had to turn to Cloudflare!

They posted an informative description of what they observed, using only a view from the "outside" of Facebook (https://blog.cloudflare.com/october-2021-facebook-outage/). Cloudflare explained that Facebook managed to withdraw BGP routes to the authoritative name servers for the facebook.com domain. Now in the DNS this would normally not be a problem, provided that the interruption to the authoritative servers is relatively short. All DNS information is cached in recursive resolvers, including name server information, and if the DNS cache time to live (TTL) is long (and by "long" I mean a day or longer) then it's likely that only a small proportion of recursive resolvers would have their cached values expire over a short (order of seconds) outage, and any user who used multiple diverse recursive resolvers would not notice the interruption at all. After all, the Facebook domain names are widely used (remember those 3 billion Facebook users?) so it is probably a widely cached name. So caching would help in a "normal" case.

At this point the second factor of this outage kicks in. Facebook uses short TTLs in their DNS, so the effect of a withdrawal of reachability of their authoritative name servers was relatively immediate. As the local cached entries timed out the authoritative servers were uncontactable, so the name disappeared from the Internet's recursive resolvers.

But this form of disappearance in the DNS is a form that raises the ire of the DNS gods. In this situation, where the name servers all go offline, then the result of a query is not an NXDOMAIN response code ("I'm

sorry but that name does not exist in the DNS, go away!") but a far more indeterminate timeout to a query with no response whatsoever. A recursive resolver will retry the query using all the name server IP addresses stored in the parent zone (.com in this case), and then return the SERVFAIL response code (which means something like: "I couldn't resolve this name, but maybe it's me, so you might want to try other resolvers before giving up!"). So, the client's stub resolver then asks the same question to all the other recursive resolvers that it has been configured with. As the Cloudflare post points out: "So now, because Facebook and their sites are so big, we have DNS resolvers worldwide handling 30x more queries than usual and potentially causing latency and timeout issues to other platforms."

Then the third factor kicked in. Once the domain name facebook.com and all the names in this space effectively disappeared from the Internet then their own internal command and control tools also disappeared. Whether this was a consequence of the DNS issue, or a consequence of the original BGP route withdrawal isn't possible to determine from here, but the result was that they lost control of the service platform. And this then impacted the ability for their various data centres to exchange traffic, which further exacerbated the problem. As Facebook's note admitted, the outage "impacted many of the internal tools and systems we use in our day-to-day operations, complicating our attempts to quickly diagnose and resolve the problem." Other reports in Twitter were more fanciful, including a report that the Facebook office buildings defaulted to a locked mode, preventing staff from entering the facilities, presumably in order to work on the outage! At this point Facebook's service platform was presumably running solo, as no one could get into the platform and chase down the issue and rectify it directly. They were evidently locked out!

There are numerous lessons to learn from this outage, so let's look at a few:

- Rehearse every config change, and always have a plan to back out. Need I say more?

- TTL values in the DNS. If you want to use short TTLs on your DNS then tread very carefully, because DNS caching will not be there to help you get out of a jam. On the whole, DNS caching is what makes the Internet work so efficiently. Reducing cache lifetimes simply pushes you ever closer to the edge of disaster!

- Don't put all your DNS eggs in one basket. The DNS will cooperate quite happily with diverse redundancy, but you need to set this up to make it work for you. So don't place all your authoritative DNS name servers in a single AS in BGP. That's just asking for trouble. Maybe you want to enlist the help of one or two external providers to act as an additional secondary server for your domain.

- It's always good to split out your command-and-control plane from the production service. That way you will always be able to access your service elements even in the event of a service failure. This means separate infrastructure, separate routes, and separate DNS domains. Separate everything. Yes, there is a cost to set up an Out Of Band service that you desperately hope you will never need to rely on. But Facebook is big. I'm sure they can afford it!

- Moving fast and breaking things only ends up breaking things. At some point users lose patience, and once the users have deserted you, then you can move as fast as you like and break as much as you want. Without any customers it just doesn't matter anymore!

- If you really must make a design decision between speed and resiliency, then think very carefully about what risk you are willing to take and the downside of failure. In this case Facebook's failure has meant some 6% being wiped off the market capitalisation of the company (Oh yes, the testimony in front of the US Senate Commerce Subcommittee on Consumer Protection around the same time as the outage didn't help either). But at some point, there is an engineering trade-off between the costs of additional resiliency measures and the cost of failure of critical single points of vulnerability that can cascade. Now

> Facebook may have deliberately chosen a high-risk profile, but is this risk profile that one that necessarily is shared by its consumers, advertisers, and stockholders?

If you want your customers, your investors, your regulators, and the broader community to have confidence in you and have some assurance that you are doing are doing an effective job, then you need to be open and honest about what you are doing and why. The entire structure of public corporate entities was intended to reinforce that assurance by insisting on full and frank public disclosure of the corporate's actions. However, these are not rules that seem to apply to Facebook.

Now I could be surprised if, in the coming days, Facebook released a more comprehensive analysis of this outage, including root cause analysis, and the factors that lead to cascading failures. It could explain why efforts to rectify the immediate failure took an amazingly long 6 hours. It could describe the measures they took to restore their service and their longer-term actions that they will undertake to avoid similar failure scenarios in the future. It could detail the risk profile that guides their engineering design decisions and how this affects service resilience. And more.

Yes, I could be surprised if this were to happen.

But, between you and me, I have absolutely no such expectations. And, I suspect, neither do you!

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre.

Author

Geoff Huston AM, B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region.

www.potaroo.net