The ISP Column
A column on various things Internet

July 2021
Geoff Huston

Outage Reporting

Much has been said about the criticality of the small coterie of large-scale content distribution platforms and their critical role in today's Internet. These days when one of the small set of core content platforms experiences a service outage then it's mainstream news, as we saw in June of this year with outages reported in both Fastly and Akamai (https://www.potaroo.net/ispcol/2021-07/cdn.html). In the case of Akamai, the June outage impacted three of Australia's largest banks, their national postal service, the country's reserve bank, and one airline operator. Further afield from Australia the outage impacted the Hong Kong Stock Exchange, and some US airlines. The roll call of impacted services appeared to reach some 500 services from this Akamai incident. With Fastly's outage earlier in the month we saw a set of popular services disappear for an hour or more. The list of impacted services included Twitch, Pinterest, Reddit, Spotify, the New York Times, and the BBC, to name just a few. And now at the end of July Akamai have managed to do it again on a grander scale.

I've already talked about the increasing criticality of Cloud and Content service providers and the vulnerabilities associated with the strong levels of provider aggregation in this space. With so many enterprises all over the Internet forced to make a choice between just a handful of viable content distribution platforms for their content and services then nobody should be surprised when a single platform's outage has massive service impact. But that's not what's prompted me to write this note.

Akamai's report of the incident was unusual. I'll reproduce here in full:

> [07:35 UTC on July 24, 2021] Update:
>
> Root Cause:
>
> > This configuration directive was sent as part of preparation for independent load balancing control of a forthcoming product. Updates to the configuration directive for this load balancing component have routinely been made on approximately a weekly basis. (Further changes to this configuration channel have been blocked until additional safety measures have been implemented, as noted in Corrective and Preventive Actions.)
> >
> > The load balancing configuration directive included a formatting error. As a safety measure, the load balancing component disregarded the improper configuration and fell back to a minimal configuration. In this minimal state, based on a VIP-only configuration, it did not support load balancing for Enhanced TLS slots greater than 6145.
> >
> > The missing load balancing data meant that the Akamai authoritative DNS system for the akamaiedge.net zone would not receive any directive for how to respond to DNS queries for many Enhanced TLS slots. The authoritative DNS system will respond with a SERVFAIL when there is no directive, as during localized failures resolvers will retry an alternate authority.
> >
> > The zoning process used for deploying configuration changes to the network includes an alert check for potential issues caused by the configuration changes. The zoning process did result in alerts during the deployment. However, due to how the particular safety check was configured, the alerts for this load balancing component did not prevent the configuration from continuing to propagate, and did not result in escalation to engineering SMEs. The input safety check on the load balancing component also did not automatically roll back the change upon detecting the error.
>
> Contributing Factors:
>
> > The internal alerting which was specific to the load balancing component did not result in blocking the configuration from propagating to the network, and did not result in an escalation to the SMEs for the component.

>> The alert and associated procedure indicating widespread SERVFAILs potentially due to issues with mapping systems did not lead to an appropriately urgent and timely response.
>>
>> The internal alerting which fired and was escalated to SMEs was for a separate component which uses the load balancing data. This internal alerting initially fired for the Edge DNS system rather than the mapping system, which delayed troubleshooting potential issues with the mapping system and the load balancing component which had the configuration change. Subsequent internal alerts more clearly indicated an issue with the mapping system.
>>
>> The impact to the Enhanced TLS service affected Akamai staff access to internal tools and websites, which delayed escalation of alerts, troubleshooting, and especially initiation of the incident process.

> Short Term
>
> Completed:

>> Akamai completed rolling back the configuration change at 16:44 UTC on July 22, 2021.
>>
>> Blocked any further changes to the involved configuration channel.
>>
>> Other related channels are being reviewed and may be subject to a similar block as reviews take place. Channels will be unblocked after additional safety measures are assessed and implemented where needed.

> In Progress:

>> Validate and strengthen the safety checks for the configuration deployment zoning process
>>
>> Increase the sensitivity and priority of alerting for high rates of SERVFAILs.

> Long Term
>
> In Progress:

>> Reviewing and improving input safety checks for mapping components.
>>
>> Auditing critical systems to identify gaps in monitoring and alerting, then closing unacceptable gaps.

Why do I find this report unusual?

It’s informative in detailing their understanding of the root cause of the problem, 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.

I guess we've become used reading evasive and vague outage reports that talk about "operational anomalies" causing "service incidents" that are "being rectified by our stalwart team of engineers as we speak". When we see a report that details the issues and the remedial measures it sticks out as a welcome deviation from the mean. It’s as if any admission of the details of a fault in the service exposes the provider to some form of ill-defined liability or reputation damage, and to minimise this exposure the reports of faults, root causes and mediation actions are all phrased in terms of vague and meaningless generalities.

Other industries have got over this defensive stance, albeit in some cases with a little outside assistance. The airline industry is a good case in point where the intent of such investigations is not to attribute blame and determine liability, but to determine the causes of the incident and understand how such circumstances can be avoided in the future because of the obvious overarching safety concerns. Other industries, including the automobile industry, the nuclear power industry, the chemical industry have all been taught the sometimes-painful lesson that the path to a safer service and safer products necessarily involves an open, dispassionate, and honest investigation into incidents with the service. Incidents are an opportunity to learn why a system fails, and an honest and comprehensive post-event analysis can offer invaluable pointers as to what measures can be taken to avoid similar failure modes in the future. It allows all service providers to operate a safer service.

Yet despite these practices that have been adopted in other industries, the information technology industry often regards itself as "special”. For decades software vendors have been able to sell faulty and insecure product without even a hint of liability, and the effort to improve the robustness of the product was often seen as an avoidable cost to the software vendor. This attitude is still pervasive in this industry and manifests itself in outages on the Internet with depressing regularity. "Move fast and break things" became a pervasive mantra of the Internet, and not only did Facebook's Mark Zuckerberg adopt this as the operating principle for Facebook's internal engineering effort, but he went further to observe that "Unless you are breaking stuff, you are not moving fast enough." Perhaps we should simply be grateful that Facebook does not build aeroplanes, nuclear power plants or automobiles. But this mantra of rapid and at times somewhat careless innovation isn't unique to Facebook, and has been applicable equally to many others, including Amazon, Apple, and Google, who have all been moving extremely quickly and doubtless they all have been breaking a few things along the way!

This industry's case for special immunity from such dispassionate and thorough investigations into failures and incidents may well have been based on an assumption that failures of these systems were not immediately and directly related to public safety. Yet this is simply not true. A failure in the ability of a DNS platform to resolve names may sound like a relatively obscure and inconsequential failure, but if other systems rely on the DNS in quite fundamental ways, then we head into a cascading failure scenario. The more we use digital systems as the command-and-control mechanism for our public environment, such as our power supply systems, oil pipelines, transportation networks, and water supply systems to name just a few, then the more we implicitly rely on the safe and robust operation of basic digital infrastructure services. It's not just the waves of various cyber-attacks that expose such vulnerabilities in our digital world, but the issues of failures in these highly complex systems where supposedly minor changes to the system or its environment can lead to catastrophic failure either through unintended consequences or more troublesome aspects of emergent behaviours of highly complex systems.

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 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.

Disclaimer

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

Author

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

www.potaroo.net