the Availability Digest

www.availabilitydigest.com
@availabilitydig

Let's Share Outage Information for the Benefit of All

Andrew Gallo
Senior Information Systems Engineer
April 2014

"It's impossible to solve significant problems using the same level of knowledge that created them!" –Albert Einstein

Outages happen - it's a simple fact of running any type of system, be it network, server, application, aviation, nuclear, etc. Urs Hölzle, a Distinguished Fellow at Google and it's first vice president of engineering, plainly states it this way- "at scale, everything breaks." Outages are disruptive, potentially dangerous, and sometimes embarrassing. Much of what we do as engineers is to design systems to avoid or minimize the chance and impact of these outages. Our vendors and standards bodies design high availability technologies such as VMWare's Fault Tolerance or vMotion, the IETF's VRRP, RAID, etc. This article is not about the merits of any of these technologies, how to deploy them, or what to do to recover when they fail. Rather, this is a call to share our lessons learned after an analysis of what the failure was and how it could have been avoided. The tough lessons learned in these situations are too valuable to be kept secret.[1]

What Makes a Good Outage Report?

This isn't a tutorial on why you should write a post mortem or who should be responsible for one. Let's focus on the content, the "what." Plus, there are several methodologies. Having said that, a good report should contain a detailed description of the event, how it was recognized (e.g., customer reports or monitoring systems), the scope and severity of the outage, and what was done to restore service, both temporary (if appropriate) and permanent fixes. A timeline of events leading up to and during the outage as well as during restoration should be provided. Personnel involved and key decisions (along with their justifications and if they were the right decisions) should be included.

The summary should contain a direct cause analysis, contributing cause analysis, and a root cause analysis. Generally, these are defined as:

- Direct cause – what lead directly or immediately to the occurrence
- Contributing cause – causes or factors that, by themselves, would not have caused a problem, but when present, either allowed or worsened the problem
- Root cause – the underlying conditions that lead to the outage. If these causes weren't present, the outage would not have happened. Some believe root causes must always be traced back to human causes or failings, whereas the other two can be purely technical.

Additionally, the summary should include remediation of these causes so that this outage will not be repeated in the future (at least by the same causes).

[1] This article was first published as a blog by Andrew Gallo on April 2, 2014 in Packet Pushers.
http://packetpushers.net/outages-suck-lets-share-action-reports/?utm_campaign=blogs&utm_medium=twitter&utm_source=twitter

Why We Don't Share

While it's fairly obvious why we don't analyze or share such experiences, let's state them for completeness. One of the reasons we don't thoroughly analyze and then share our outage experiences is the lack of time and resources. We need to get systems up and operational, so documentation is a luxury. Self-reflection also requires maturity in an organization. The immediate operational need to restore service, resolve the problem, and move on is often more pressing than analyzing, documenting, and sharing your findings.

Running from one fire to the next, one project to the next, without building a knowledge base of lessons learned is a sign of a distressed IT shop. This is an organizational problem and we should lead our organizations to mature to a point where critical self-analysis is part of our routine in order to better ourselves.

Second, it's embarrassing. Many of us take pride in the systems and applications we build, and when they fail we might take it as a sign of personal and professional failure. A little humility goes a long way to being a confident professional. I would suggest that sharing such information might be a way to combat professional loneliness.

Then there are the barriers of security, competition, and liability. A well sanitized report focusing on the technology goes a long way to address the first two reasons, though there may be times when certain information can't be shared. As for the legal issue, that's tougher, though there are examples that show how this can work in certain cases.

Examples for Formalized Incident Reporting Systems

There are several examples of formalized incident reports or regular places to share incident reports, though sadly, most of them are outside of the information & communications industry.

Bob "@strat" Stratton gave an excellent FireTalk at ShmooCon 2014 discussing information sharing, primarily focused on IT security, by using the Aviation Safety Reporting System as a model that should be followed. The Aviation Safety Reporting System (ASRS) is a voluntary, non-punitive system of reporting safety and other problems in aviation for the purpose of analysis and dissemination of vital safety information to the community. According to the presentation, this function was originally under the Federal Aviation Administration (FAA), but because this agency had enforcement powers, people were not volunteering information. To encourage more reporting and to protect those making reports, this function was moved to the National Aeronautics and Space Administration (NASA) with explicit legal prohibition for using information in these reports for enforcement action (see 14 CFR §91.25).

Still in the aviation arena, Flying Magazine has a long-standing column entitled "I Learned About Flying From That," where (generally private) pilots can share incidents with their peers in the hopes that the hard and potentially life threatening lessons learned by one individual don't need to be repeated.

More broadly with transportation, the National Transportation Safety Board (NTSB) publishes voluminous evidence, investigation notes, and findings of its thorough and scientific investigations of transportation accidents.

The Nuclear Regulatory Commission (NRC) has an Event Notification reporting system that requires licensees to notify the commission of any number of unusual events or emergencies such as misadministration of radio isotopes for medical treatment, nuclear power reactor trips or SCRAMs, loss or inability to locate exit signs that use tritium continuous lighting without power , and showing up to work drunk or high.

The regulated telecommunications industry has a similar system, the Federal Communications Commission's Network Outage Reporting System (NORS). Sadly, and unlike the NRC's reports, these are not open for public access. The previous version of this system, the Wireline Outage Reports. focused almost exclusively on local and long distance wireline voice providers. It had been open to the public, but for "security" reasons it is now hidden. 47 CFR Part 4 lays out the requirements of what types of services and organizations are subject to reporting along with the thresholds that triggers a report. Section 4.2 is what provides the confidentiality, which appears in the regulations starting in 2005. Providers such as AT&T commented on this section stating "The comments establish overwhelmingly that outage reports should not be made available to the public because such disclosure would create grave risks to the Nation's critical infrastructure security." (page 8 in this pdf) One wonders what the justification for this action was given the NRC's reports remaining open.

Early in my career, I remember reading these and being fascinated and a bit confused, trying to figure out how the PSTN worked. There were many fascinating reports, including this one, luckily archived, where Sprint filed a report on a special circuit to Area 51.

What We Can and Should Be Doing

First, within our organizations, we should have a regular practice of after-incident evaluation and documentation for purposes of improving service availability. Investigations and findings should be open, honest, shared, and where appropriate, remediations implemented.

Second, we should share these findings and experiences with our colleagues outside our organizations. As mentioned above, some industries have formalized (and required) reporting mechanisms. Short of that, blogs such as this one, mailing lists such as Outages with its associated wiki page (though these are primarily concerned with planned & unplanned outage notification), or the North American Network Operators Group might be appropriate places to share such post incident experiences. The Availability Digest (@AvailabilityDig) is also a good resource for sharing articles on failures and lessons learned.

I'll go back to the nuclear industry, which seems more open than our own, to quote from a report written for the The National Diet of Japan, "The official report of The Fukushima Nuclear Accident Independent Investigation Commission:"

> "A global perspective should be emphasized, so that the results and conclusions will help to prevent nuclear accidents elsewhere."

This is an extreme case of a deadly accident with the output shared with the world to avoid its repetition.

Maybe I'm being a bit naive in thinking that we can look past embarrassment, competition, and liability in order to challenge and teach each other to be better engineers.

Andrew Gallo
Senior Information Systems Engineer

Andrew Gallo is a Washington, DC based Senior Information Systems Engineer and Network Architect responsible for design and implementation of the enterprise network for a large university. Areas of specialization include the University's wide area connections, including a 150 kilometer DWDM ring, designing a multicampus routing policy, and business continuity planning for two online datacenters. Andrew started during the internet upswing of the mid to late 90s installing and terminating fiber. As his career progressed, he has had experience with technologies from FDDI to ATM, and all speeds of Ethernet, including a recent deployment of several metro-area 100Gbps circuits. Focusing not only on data networks, Andrew has experience in traditional TDM voice,

VoIP, and real-time, unified collaboration technologies. Areas of interest include optical transport, network virtualization and software defined networking, and network science and graph theory.