Chapter 15 - Postmortem Culture: Learning from Failure

Postmortem Culture: Learning from Failure

Written by John Lunney and Sue Lueder
Edited by Gary O' Connor

> The cost of failure is education.
>
> Devin Carraway

As SREs, we work with large-scale, complex, distributed systems. We constantly enhance our services with new features and add new systems. Incidents and outages are inevitable given our scale and velocity of change. When an incident occurs, we fix the underlying issue, and services return to their normal operating conditions. Unless we have some formalized process of learning from these incidents in place, they may recur ad infinitum. Left unchecked, incidents can multiply in complexity or even cascade, overwhelming a system and its operators and ultimately impacting our users. Therefore, postmortems are an essential tool for SRE.

The postmortem concept is well known in the technology industry [All12]. A postmortem is a written record of an incident, its impact, the actions taken to mitigate or resolve it, the root cause(s), and the follow-up actions to prevent the incident from recurring. This chapter describes criteria for deciding when to conduct postmortems, some best practices around postmortems, and advice on how to cultivate a postmortem culture based on the experience we've gained over the years.

Google's Postmortem Philosophy

The primary goals of writing a postmortem are to ensure that the incident is documented, that all contributing root cause(s) are well understood, and, especially, that effective preventive actions are put in place to reduce the likelihood and/or impact of recurrence. A detailed survey of root-cause analysis techniques is beyond the scope of this chapter (instead, see [Roo04]); however, articles, best practices, and tools abound in the system quality domain. Our teams use a variety of techniques for root-cause analysis and choose the technique best suited to their services. Postmortems are expected after any significant undesirable event. Writing a postmortem is not punishment—it is a learning opportunity for the entire company. The postmortem process does present an inherent cost in terms of time or effort, so we are deliberate in choosing when to write one. Teams have some internal flexibility, but common postmortem triggers include:

- User-visible downtime or degradation beyond a certain threshold
- Data loss of any kind
- On-call engineer intervention (release rollback, rerouting of traffic, etc.)
- A resolution time above some threshold
- A monitoring failure (which usually implies manual incident discovery)

It is important to define postmortem criteria before an incident occurs so that everyone knows when a postmortem is necessary. In addition to these objective triggers, any stakeholder may request a postmortem for an event.

Blameless postmortems are a tenet of SRE culture. For a postmortem to be truly blameless, it must focus on identifying the contributing causes of the incident without indicting any individual or team for bad or inappropriate behavior. A blamelessly written postmortem assumes that everyone involved in an incident had good intentions and did the right thing with the information they had. If a culture of finger pointing and shaming individuals or teams for doing the "wrong" thing prevails, people will not bring issues to light for fear of punishment.

Blameless culture originated in the healthcare and avionics industries where mistakes can be fatal. These industries nurture an environment where every "mistake" is seen as an opportunity to strengthen the system. When postmortems shift from allocating blame to investigating the systematic reasons why an individual or team had incomplete or incorrect information, effective prevention plans can be put in place. You can't "fix" people, but you can fix systems and processes to better support people making the right choices when designing and maintaining complex systems.

When an outage does occur, a postmortem is not written as a formality to be forgotten. Instead the postmortem is seen by engineers as an opportunity not only to fix a weakness, but to make Google more resilient as a whole. While a blameless postmortem doesn't simply vent frustration by pointing fingers, it *should* call out where and how services can be improved. Here are two examples:

Pointing fingers

> "We need to rewrite the entire complicated backend system! It's been breaking weekly for the last three quarters and I'm sure we're all tired of fixing things onesy-twosy. Seriously, if I get paged one more time I'll rewrite it myself…"

Blameless

> "An action item to rewrite the entire backend system might actually prevent these annoying pages from continuing to happen, and the maintenance manual for this version is quite long and really difficult to be fully trained up on. I'm sure our future on-callers will thank us!"

Best Practice: Avoid Blame and Keep It Constructive

Blameless postmortems can be challenging to write, because the postmortem format clearly identifies the actions that led to the incident. Removing blame from a postmortem gives people the confidence to escalate issues without fear. It is also important not to stigmatize frequent production of postmortems by a person or team. An atmosphere of blame risks creating a culture in which incidents and issues are swept under the rug, leading to greater risk for the organization [Boy13].

Collaborate and Share Knowledge

We value collaboration, and the postmortem process is no exception. The postmortem workflow includes collaboration and knowledge-sharing at every stage.

Our postmortem documents are Google Docs, with an in-house template (see Example Postmortem). Regardless of the specific tool you use, look for the following key features:

Real-time collaboration

> Enables the rapid collection of data and ideas. Essential during the early creation of a postmortem.

An open commenting/annotation system

> Makes crowdsourcing solutions easy and improves coverage.

Email notifications

> Can be directed at collaborators within the document or used to loop in others to provide input.

Writing a postmortem also involves formal review and publication. In practice, teams share the first postmortem draft internally and solicit a group of senior engineers to assess the draft for completeness.

Review criteria might include:

- Was key incident data collected for posterity?
- Are the impact assessments complete?
- Was the root cause sufficiently deep?
- Is the action plan appropriate and are resulting bug fixes at appropriate priority?
- Did we share the outcome with relevant stakeholders?

Once the initial review is complete, the postmortem is shared more broadly, typically with the larger engineering team or on an internal mailing list. Our goal is to share postmortems to the widest possible audience that would benefit from the knowledge or lessons imparted. Google has stringent rules around access to any piece of information that might identify an end-user,[80] and even internal documents like postmortems never include such information.

Best Practice: No Postmortem Left Unreviewed

An unreviewed postmortem might as well never have existed. To ensure that each completed draft is reviewed, we encourage regular review sessions for postmortems. In these meetings, it is important to close out any ongoing discussions and comments, to capture ideas, and to finalize the state.

Once those involved are satisfied with the document and its action items, the postmortem is added to a team or organization repository of past incidents.[81] Transparent sharing makes it easier for others to find and learn from the postmortem.

Introducing a Postmortem Culture

Introducing a postmortem culture to your organization is easier said than done; such an effort requires continuous cultivation and reinforcement. We reinforce a collaborative postmortem culture through senior management's active participation in the review and collaboration process. Management can encourage this culture, but blameless postmortems are ideally the product of engineer self-motivation. In the spirit of nurturing the postmortem culture, SREs proactively create activities that disseminate what we learn about system infrastructure. Some example activities include:

Postmortem of the month

> In a monthly newsletter, an interesting and well-written postmortem is shared with the entire organization.

Google+ postmortem group

> This group shares and discusses internal and external postmortems, best practices, and commentary about postmortems.

Postmortem reading clubs

> Teams host regular postmortem reading clubs, in which an interesting or impactful postmortem is brought to the table (along with some tasty refreshments) for an open dialogue with participants, nonparticipants, and new Googlers about what happened, what lessons the incident imparted, and the aftermath of the incident. Often, the postmortem being reviewed is months or years old!

Wheel of Misfortune

> New SREs are often treated to the Wheel of Misfortune exercise (see Disaster Role Playing), in

which a previous postmortem is reenacted with a cast of engineers playing roles as laid out in the postmortem. The original incident commander attends to help make the experience as "real" as possible.

One of the biggest challenges of introducing postmortems to an organization is that some may question their value given the cost of their preparation. The following strategies can help in facing this challenge:

- Ease postmortems into the workflow. A trial period with several complete and successful postmortems may help prove their value, in addition to helping to identify what criteria should initiate a postmortem.
- Make sure that writing effective postmortems is a rewarded and celebrated practice, both publicly through the social methods mentioned earlier, and through individual and team performance management.
- Encourage senior leadership's acknowledgment and participation. Even Larry Page talks about the high value of postmortems!

Best Practice: Visibly Reward People for Doing the Right Thing

Google's founders Larry Page and Sergey Brin host TGIF, a weekly all-hands held live at our headquarters in Mountain View, California, and broadcast to Google offices around the world. A 2014 TGIF focused on "The Art of the Postmortem," which featured SRE discussion of high-impact incidents. One SRE discussed a release he had recently pushed; despite thorough testing, an unexpected interaction inadvertently took down a critical service for four minutes. The incident only lasted four minutes because the SRE had the presence of mind to roll back the change immediately, averting a much longer and larger-scale outage. Not only did this engineer receive two peer bonuses[82] immediately afterward in recognition of his quick and level-headed handling of the incident, but he also received a huge round of applause from the TGIF audience, which included the company's founders and an audience of Googlers numbering in the thousands. In addition to such a visible forum, Google has an array of internal social networks that drive peer praise toward well-written postmortems and exceptional incident handling. This is one example of many where recognition of these contributions comes from peers, CEOs, and everyone in between.[83]

Best Practice: Ask for Feedback on Postmortem Effectiveness

At Google, we strive to address problems as they arise and share innovations internally. We regularly survey our teams on how the postmortem process is supporting their goals and how the process might be improved. We ask questions such as: Is the culture supporting your work? Does writing a postmortem entail too much toil (see Eliminating Toil)? What best practices does your team recommend for other teams? What kinds of tools would you like to see developed? The survey results give the SREs in the trenches the opportunity to ask for improvements that will increase the effectiveness of the postmortem culture.

Beyond the operational aspects of incident management and follow-up, postmortem practice has been woven into the culture at Google: it's now a cultural norm that any significant incident is followed by a comprehensive postmortem.

Conclusion and Ongoing Improvements

We can say with confidence that thanks to our continuous investment in cultivating a postmortem culture, Google weathers fewer outages and fosters a better user experience. Our "Postmortems at Google" working group is one example of our commitment to the culture of blameless postmortems. This group coordinates postmortem efforts across the company: pulling together postmortem templates, automating postmortem creation with data from tools used during an incident, and helping automate data extraction

from postmortems so we can perform trend analysis. We've been able to collaborate on best practices from products as disparate as YouTube, Google Fiber, Gmail, Google Cloud, AdWords, and Google Maps. While these products are quite diverse, they all conduct postmortems with the universal goal of learning from our darkest hours.

With a large number of postmortems produced each month across Google, tools to aggregate postmortems are becoming more and more useful. These tools help us identify common themes and areas for improvement across product boundaries. To facilitate comprehension and automated analysis, we have recently enhanced our postmortem template (see Example Postmortem) with additional metadata fields. Future work in this domain includes machine learning to help predict our weaknesses, facilitate real-time incident investigation, and reduce duplicate incidents.

[80] See *http://www.google.com/policies/privacy/*.

[81] If you'd like to start your own repository, Etsy has released *Morgue*, a tool for managing postmortems.

[82] Google's Peer Bonus program is a way for fellow Googlers to recognize colleagues for exceptional efforts and involves a token cash reward.

[83] For further discussion of this particular incident, see Emergency Response.

previous

Chapter 14 - Managing Incidents

next

Chapter 16 - Tracking Outages