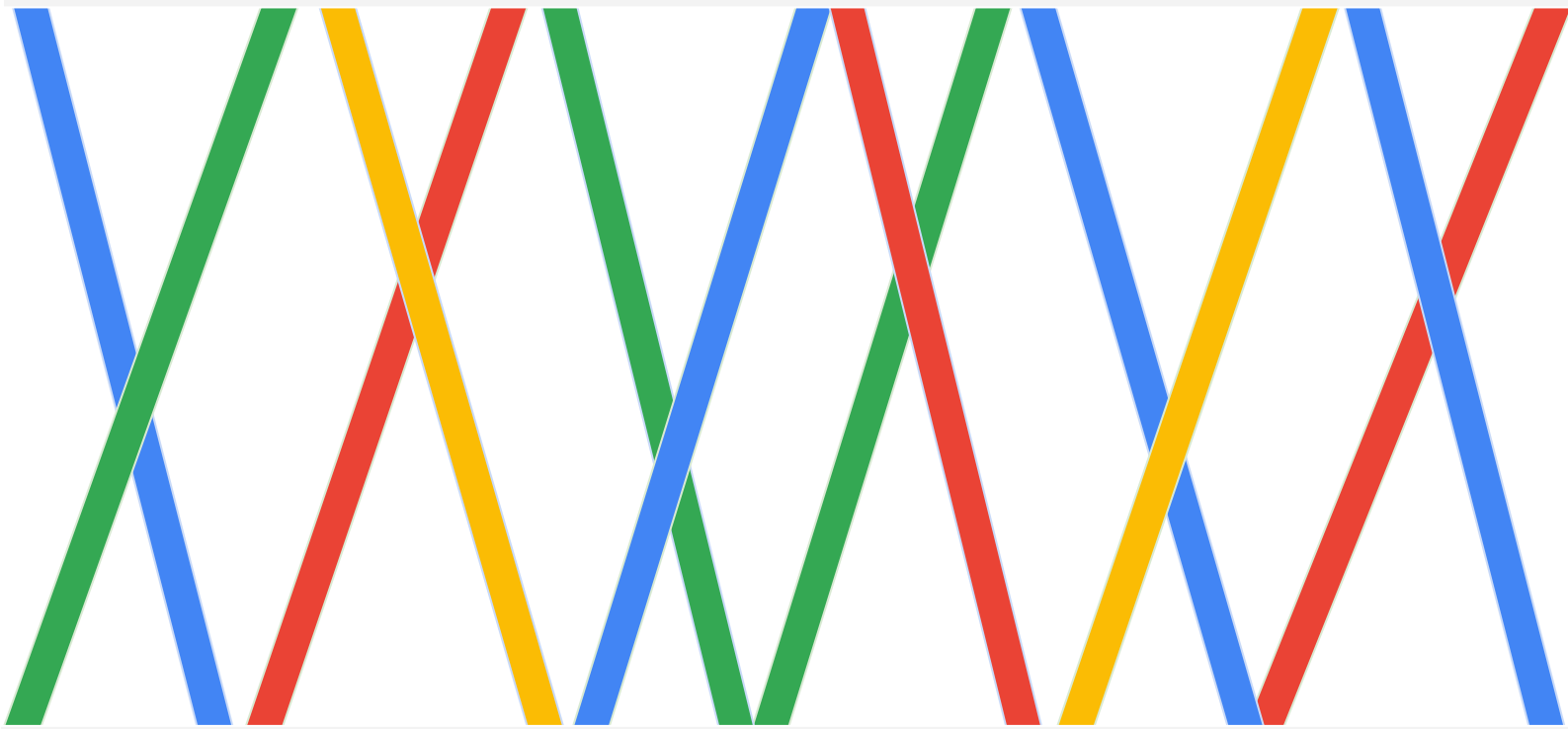
Incident Management Guide

Special edition publication commemorating

Google Site Reliability Engineering Team's 20th Anniversary



Google Site Reliability Engineering

Google
Site Reliability Engineering:
Incident Management Guide

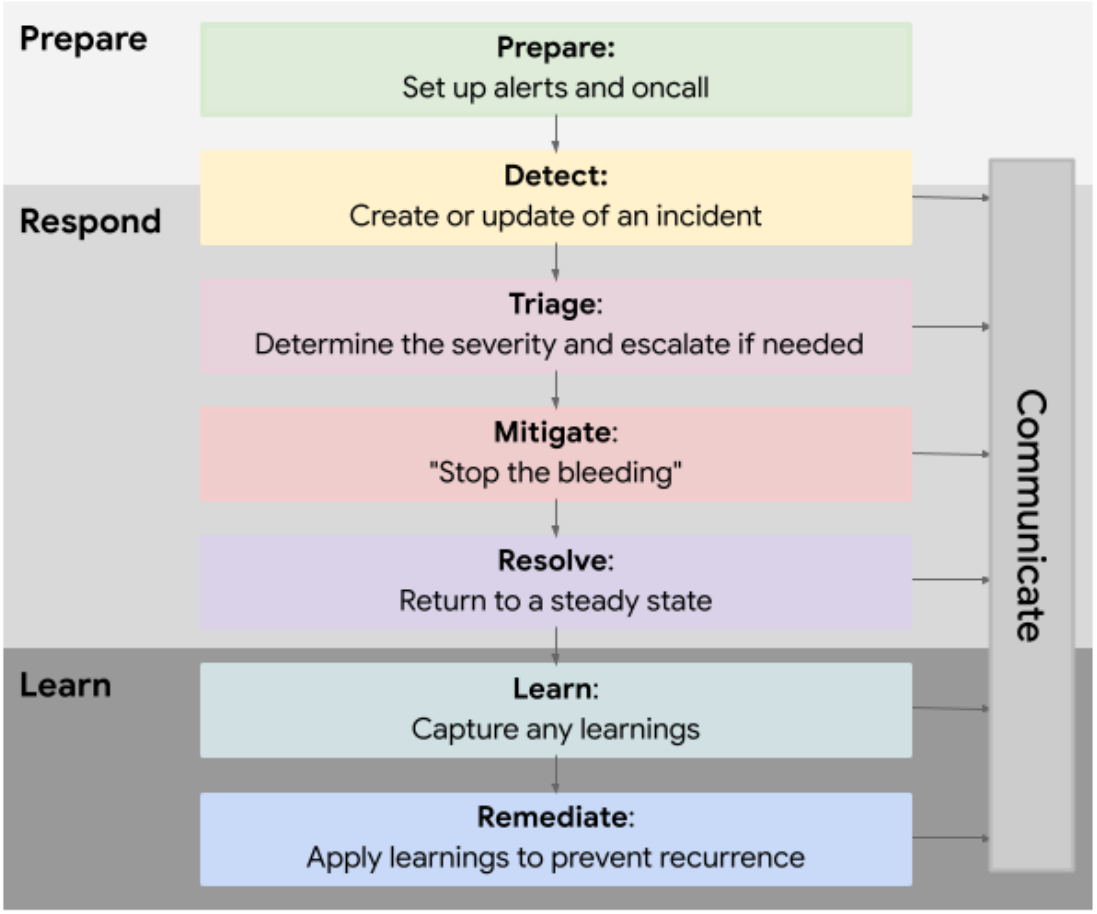
Written by

Adam Crume, Alex Cepoi, Chelsea Granados, Roxana Loza, Steve McGhee, Svetlana Gites, Trevor Mattson-Hamilton, and Vrai Stacey

Google | Site Reliability Engineering

Introduction

Outages are inevitable in any sufficiently complex system. When an outage happens, it is essential to have a process in place to know how to manage and respond effectively. A well-defined plan helps teams minimize the impact on users and customers, coordinate their response to mitigate the incident faster, and learn from it so that it can be prevented from happening again. Google has a long-standing, well-documented incident response process that has been developed and refined over the years. *Here we present a primer of the end-to-end process.*



Overview of Google's Incident Response Management

Prepare for incidents

Effective incident response begins with preparation, which includes having in place a reliable alerting mechanism and well-defined oncall process. Alerts are an efficient way to detect system issues and notify the on-call team so they can be addressed.

Here are some of the key attributes of a good alerting mechanism:

- **Alert in a timely manner**: Minimize the user impact prior to incident response beginning.
- **Cover all key user facing functionality**
- **Alert based on symptoms, not causes**: Alerts should be based on end-to-end measures of customer/client experience, not based on a system's internal behavior.
- **Be actionable**: Alerts that cannot be acted upon by an on-caller generate noise.

Alerting based on SLOs (Service Level Objectives for particular functionality) is a good way to achieve the first three attributes. Some preventive alerts based on internal metrics may be required, such as protecting against an imminent failure due to approaching a hard resource quota, as failures of this nature can cause a system to instantaneously transition from 0% failure to 100% failure. However, the general rule is to avoid alerting on a system's internal behavior as these alerts don't accurately map to user impact, and are fragile due to being closely bound to a service's implementation at the time the alert is defined.

Once you have an alerting mechanism ready, you need to ensure your oncall team is ready to respond to the alerts. As with any activity, being oncall can be made much easier with proper preparation. Having up to date playbooks with instructions on how to debug and mitigate issues can speed up incident response significantly. Note that oncallers need to be aware of the playbooks, and other training material, for it to be effective. Regular practice through activities such as "Wheel of Misfortune" exercises can keep this knowledge up to date, as well as providing an opportunity for less experienced oncallers to develop their skills in a safe environment.

Where possible, automating elements of incident response will free the oncallers to focus on problem solving. This can include automation of common tasks, automated analysis of key impact information (severity, affected services/locations, etc), root cause analysis, and intelligent suggestion of mitigating actions the oncaller can take.

Respond and manage incidents

Google's incident response system, known as IMAG, is based on the Incident Command System (ICS), a US standard for responding to emergencies, such as wildfires or earthquakes. These systems focus on the "three Cs" (3Cs) of incident management: coordinate, communicate, and control.

IMAG organizes the incident response by establishing a hierarchical structure with clear roles, tasks, and communication channels. The main roles in Google's IMAG are Incident Commander (IC), Communications Lead (CL), and Operations Lead (OL). The IC coordinates the overall incident response. The CL provides regular updates to stakeholders and acts as a point of contact for incoming communications. This allows the OL to focus on mitigating the issue, minimizing user impact, and resolving the problem. This helps balance multiple ongoing needs. As suggested by the name, these leads may delegate certain tasks to other responders. Incident roles do not follow reporting chains and instead are based on knowledge and incident context.

Good incident response, like many things, is user-centric. Fixing the problem is only part of what's needed; it's just as important to ensure that your users, stakeholders, and leaders are updated about what's affected, how bad it is, what workarounds may be possible, and when the incident may be mitigated and resolved. Communicating consistently and with an appropriate level of detail for the reader builds trust and transparency. These are just as important as technical mitigations.

Google has various Incident Response Teams (IRTs) which can also be activated for additional support during major incidents. The services provided by each IRT vary, but may include coordinating multiple team-level efforts, providing hands-on assistance, identifying and contacting teams that are (or should be) involved, gathering resources, assisting in escalations, activating other IRTs, and broad internal and/or external communications.

Effective response means treating it as a project in its own right. This includes planning ahead, deciding who needs to be involved, and documenting what has been done. Chaos will naturally prevail unless it is actively managed.

Remediate and learn from incidents

One of Google's core tenets of effective incident response is to learn from outages and improve our systems to prevent similar incidents from happening in the future. When not possible, we strive to minimize the duration and impact of unavoidable/unanticipated outages. Left unchecked, outages tend to regularly resurface and accumulate over time. This increases the operational toil for the team and can lead to expended error budgets, eroded user trust, and impacted revenue. The most effective tool we have found for achieving that is through open and blameless postmortem writing.

After the incident is resolved, a write-up of the incident is immediately started, seeking to fully understand and document how the incident unfolded, its impact, as well as things that went well or could be improved. It is important to look at a broad range of aspects of the incident response, not just at fixing the immediate problem or preventing it from recurring;

looking at effective ways to improve detection, mitigation, coordination, or communication across teams and to impacted users is equally important.

One of the core tenets of SRE's culture is that postmortems should be blameless. It's important to remember that everyone involved in the incident had good intentions. Blaming individuals for unintended consequences during the response, does not aid the learning process so instead, we focus on how we can improve our systems, procedures, and training to make them more resilient.

An honest and timely postmortem write-up reviewed by stakeholders and shared broadly with the entire organization is key to identifying the most effective corrective action items to prevent similar incidents from happening again. Once SLOs for completion of action items are agreed with stakeholders, these feed back into the team's backlog. Teams balance these action items against feature work and prioritize informed by overall reliability.

Once a postmortem writing culture is established, aggregating structured data collected across a large number of postmortems to identify trends and organizational areas needing larger investments becomes a great opportunity in a larger organization.

Further reading

- SRE Book, Chapter 9: https://sre.google/workbook/incident-response/
- "Preparing for your next incident" discussion (audio): https://www.oreilly.com/content/taming-chaos-preparing-for-your-next-incident/
- How Lowe's reduced its MTTR by over 80 percent: https://cloud.google.com/blog/products/devops-sre/how-lowes-improved-incident-response-processes-with-sre
- Shrinking the impact of production incidents: https://cloud.google.com/blog/products/devops-sre/shrinking-the-impact-of-production-incidents-using-sre-principles-cre-life-lessons
- ProdEx - Google's Production Excellence Program: https://videos.itrevolution.com/watch/762364173/