Proceedings of the EMII Members Briefing: Large Scale Incident Management at Google – Nov. 2023

Briefing Organised by the EMII Board of Directors.

For further information, please contact **Prof. Caroline McMullan** (Caroline.McMullan@DCU.ie) & **Dr. Gavin Brown** (Gavin.Brown@DCU.ie).

EMII

Google

Large Scale Incident Management at Google

Emergency Management Institute Ireland, 22/Nov/2023

sre.google • @googlesre

Rory Ward

Director of Site Reliability Engineering

Site Reliability Engineering

“SRE is what happens when you ask a software engineer to design and run operations.”

Benjamin Treynor Sloss, Vice President of 24x7 Engineering, *Google*

SRE works on many different Google products

At Google scale, something is always broken.

- 0.xx% of Google's servers are down at any point in time
- Planned maintenance can take down entire data centers
- So can unplanned natural disasters
- Launches can introduce unexpected bugs in seemingly unrelated services
- Power failures happen
- Network fibre gets cut
- Data gets corrupted
- SREs break things while trying to fix them

"Culture eats strategy for breakfast"

(Commonly attributed to Peter Drucker)

The story so far

- **What is SRE?**
  - Operations -> Engineering
  - Aligns incentives of management, development, and operations
  - Promotes development velocity while enforcing reliability

- **Fundamental concepts of SRE**
  - Error budgets
  - SLOs
  - Blameless postmortems

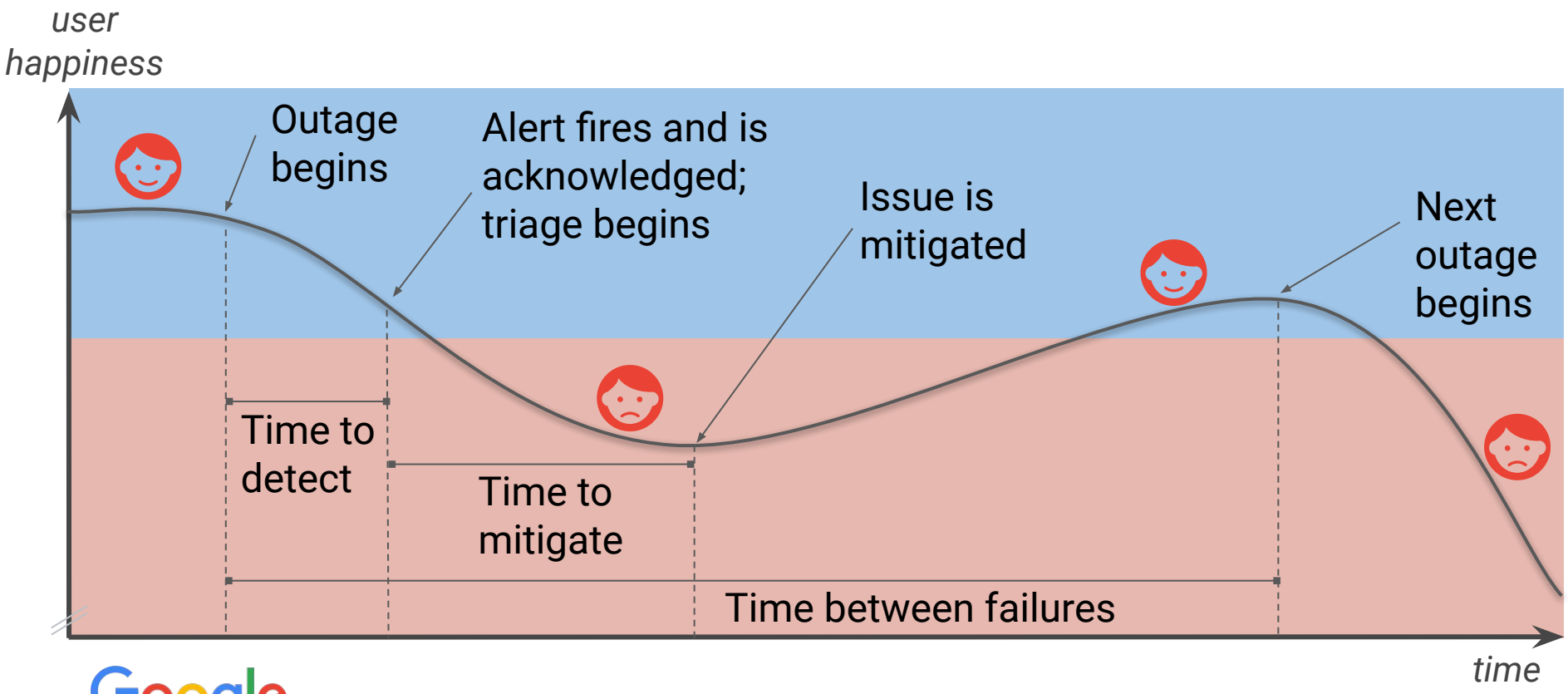
More info at https://google.com/sre

Happy users stay.

Unhappy users leave.

Google

Outage Timeline and User Happiness



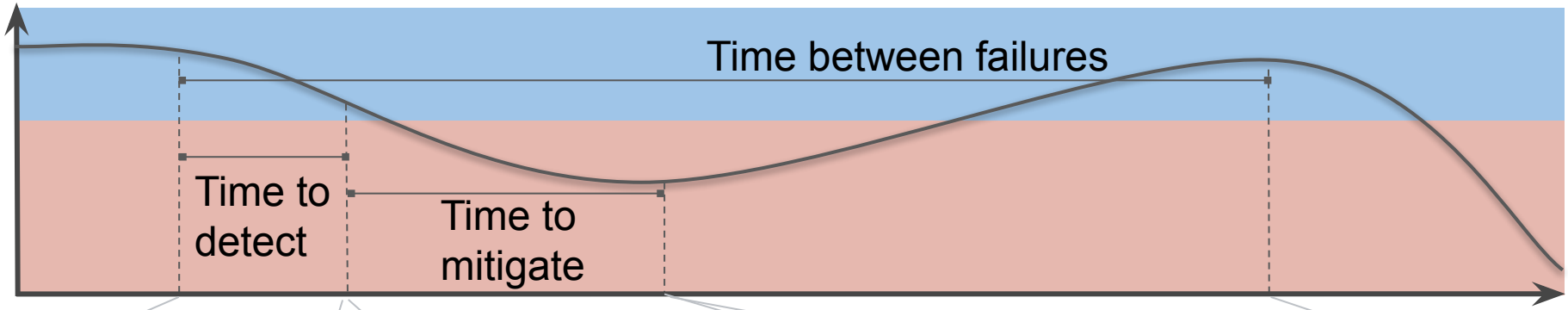
Google

Postmortem philosophy

- **The primary goals of writing a postmortem are to ensure that:**
  - The incident is documented
  - All contributing root causes are well understood
  - Effective preventive actions are put in place to reduce the likelihood and/or impact of recurrence

- **Postmortems are expected after any significant undesirable event**
  - Writing a postmortem is not a punishment

Where do we focus our effort?



- Metrics
- SLIs
- SLOs
- Alerts

- Training
- Playbooks/runbooks
- Responder fatigue
- Dashboards
- Logging

- Dev practices
- CI/CD
- Robust architectures
- Chaos engineering

Google Cloud

Service Level .*

- Service Level Indicator (**SLI**): a quantitative measure of an attribute of the service. It's a metric that users care about, such as:
  - *availability*
  - *latency*
  - *freshness*
  - *durability*
- Service Level Objective (**SLO**): SLI @ specific target
  - *99.9% target availability* = 🙂
- Service Level Agreement (**SLA**): SLO + consequences
  - *99% actual availability* = ☹️

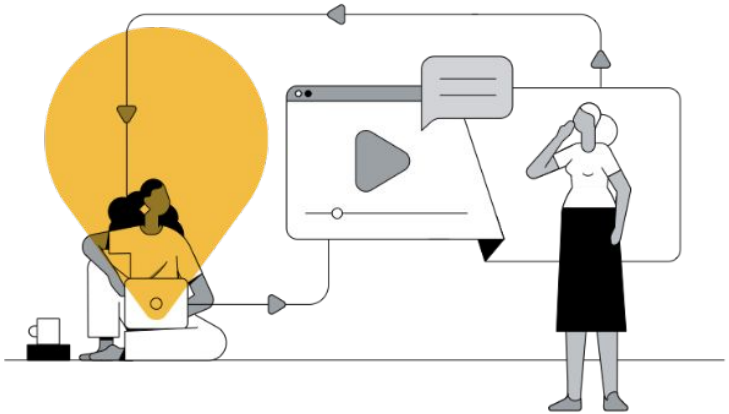
1

Intro to DiRT

Motivations

DiRT or Disaster Recovery Testing performed internally at Google is a coordinated set of events organized across the company. A group of engineers plan and execute real and fictitious outages to test the effective response of the involved teams.

Disaster Recovery Testing



Established in 2006 to exercise response to production emergencies.

Intentionally disrupt services in order to know how to respond the services and provide reliability.

It's about learning and finding single points of failure—therefore the scope of services and systems is broad.

We test software and systems, but also people, preparation, processes, and response tools.

Proprietary + Confidential

Why Google runs DiRT exercises?

SRE saying

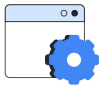
Hope is not a strategy

Report, Fix and Repeat

Vulnerabilities in systems and processes

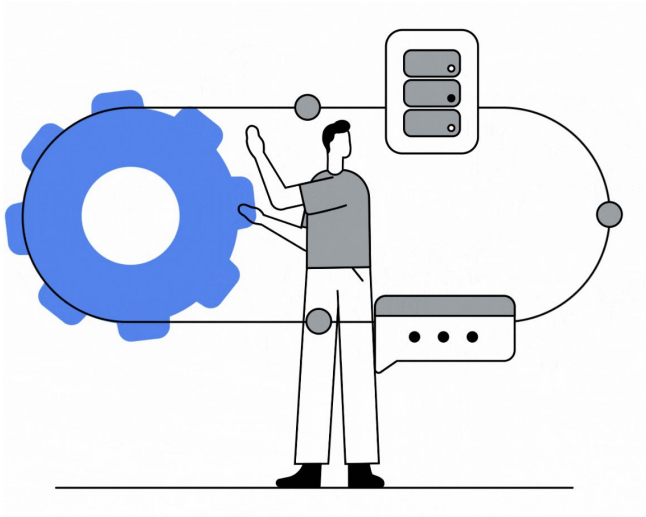
Test

Both technical and non-technical areas



Build Knowledge

Muscle memory to reduce MTTM of real incidents



Proprietary + Confidential

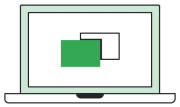
What do we Test?

Software

Modifying live service configurations, or bringing up services with known bugs

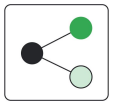
Infrastructure

Stress testing large complex architectures, validating SLOs, and ensuring resilience is maintained during disruption.



Access Controls

Including security, compliance, and privacy.



People and Workflows

Removing people who might have knowledge or experience.

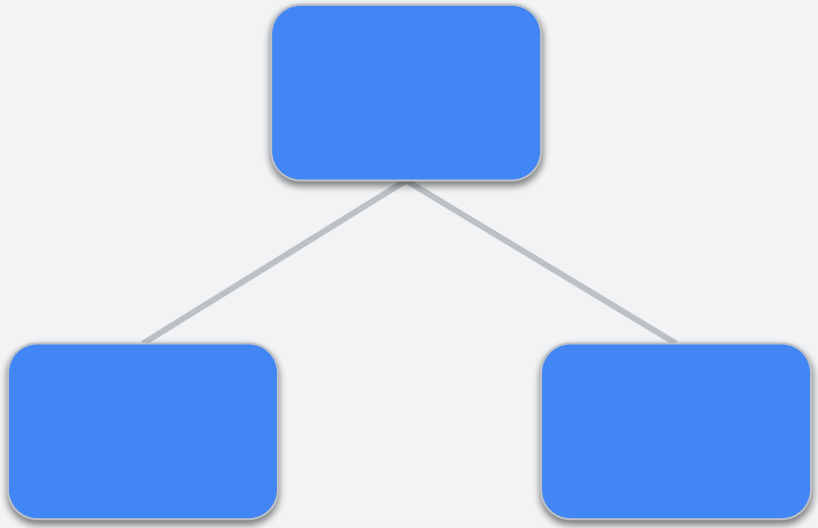
Proprietary + Confidential

Google

Incident Management At Google

WHAT is IMAG

- A familiar structure/framework/protocol, that SRE and others at Google follow when we need to respond to specific kinds of events / incidents.
- Used particularly for large incidents affecting/involving many teams.
- It is loosely modeled after the FEMA Incident Command System (ICS) used by the emergency responders in the United States.
- Assigning responsibility to different roles to support delegation.



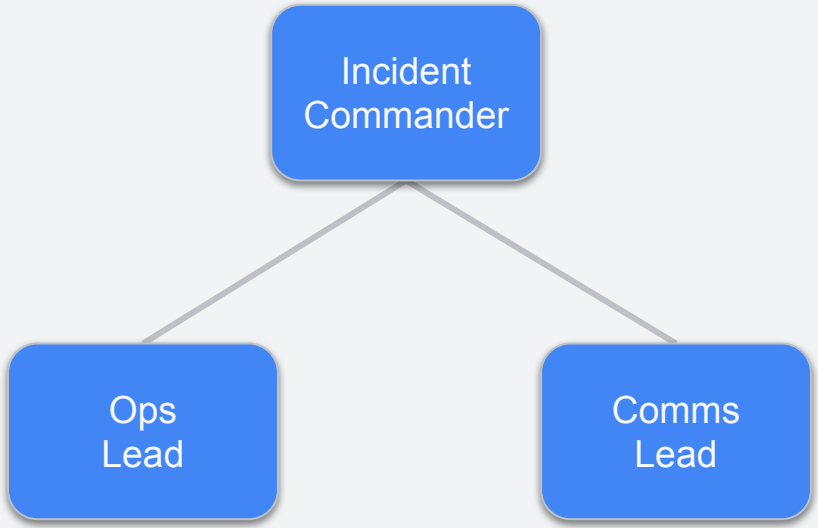
WHEN we IMAG

To handle an issue that has been ***escalated*** and requires an ***immediate***, ***continuous***, and ***organized response*** to address it.

Generic Incident Stages

What are core IMAG Roles

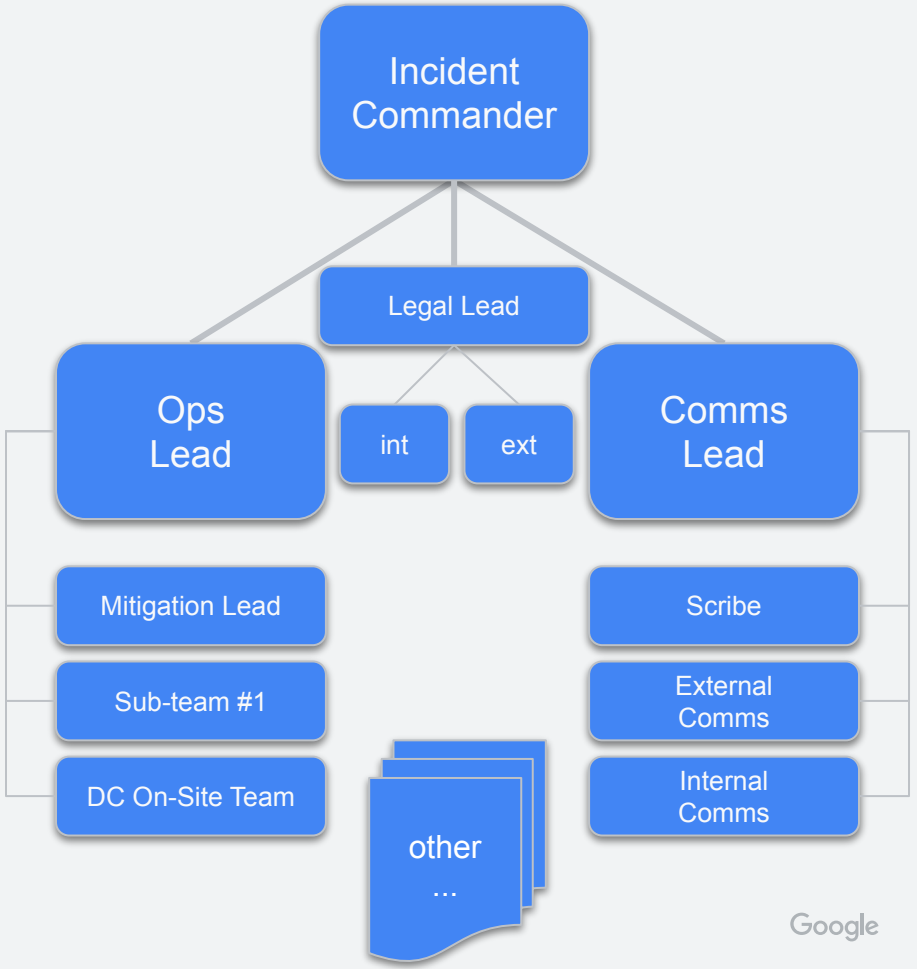
- The **Incident Commander** thinks strategically
  - A professional question-asker
  - Assigns objectives
- The **Ops Lead** is tactical
  - The do-er; How to get the objectives done
- The **Comms Lead** for communications.
  - The IC sometimes does this too.



Google

Other IMAG Roles

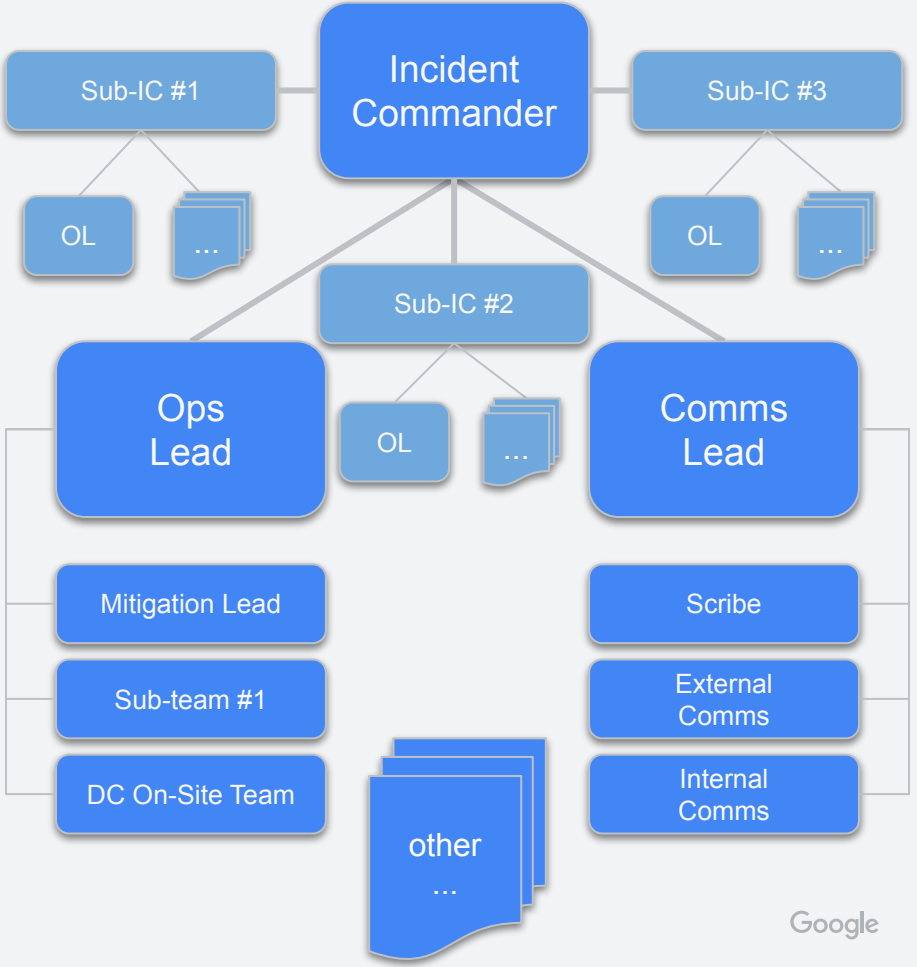
- Larger Incidents will have a more elaborate structure
- Often-seen other roles are:
  - Sub-team IC/Ops leads
  - Mitigation Leads
  - Scribe
  - External, Internal Comms
  - Legal
  - Planning Lead
- There can be others!



Google

Some Incidents are **HUGE**

- We may have multiple nested IMAG structures
- **There is still going to be one overall IC!**
- Other ICs coordinate and take direction from the overall IC



Google

WHY we IMAG

- Familiarity
  - We all use the same "language"
- Structure
  - We all know what to expect and who is who in an incident
- Scalability
  - Can address smaller and huge, planetary events
- Transferability
  - Facilitates transfer of responsibility between people and teams (handoffs)

Google

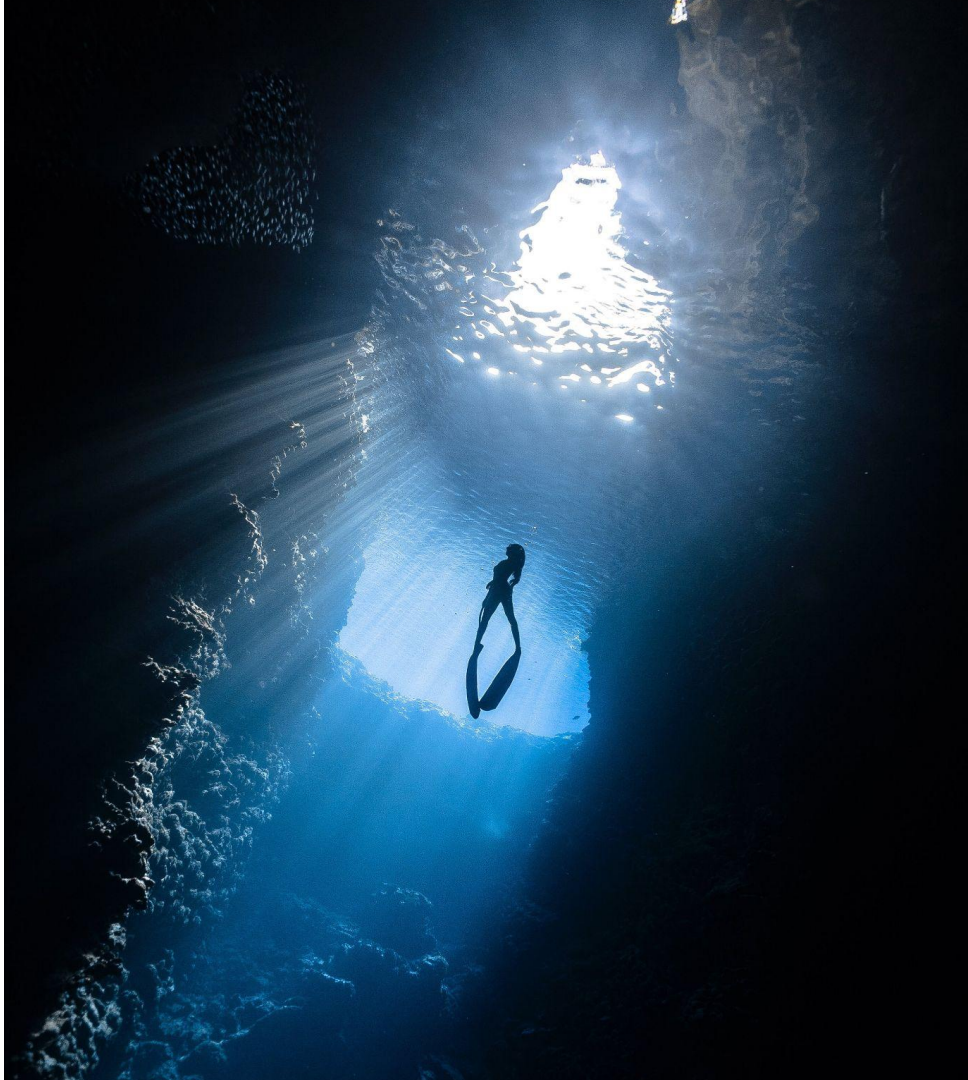
Google

Things we've learned from running production infrastructure at Google

Site Reliability Engineering

Reliability can't be taken for granted

- Easy to forget while there's plenty.
- Often too late to fix when it runs out.
- There always needs to be a voice for reliability.
- Hope is not a strategy (Google SRE motto).
- Planning for reliability needs to start early ("shift left").



Blamelessness

- Assume that everyone is competent and well-intentioned.
- Don't try "fixing" people. Fix systems and processes.
- Everyone must feel comfortable coming forward without fear of consequences.
- Only when we have the complete information, will we be able to improve.
- When problems are swept under the rug, they accumulate.

Measure what matters

- Agree on measurable goals (SLOs) to prevent conflicts.
- Focus on the user - measure what they care about.
- Anything you don't measure gets worse.
- Advise, don't block - people find ways around the gatekeeper.

No heroes

- Heroism is bad - for the hero, the team, and the system.
- SRE > SLA
- The oncaller's job is to make sure the problem is fixed, not to fix the problem.
- The oncaller is never alone. Escalate.

Change is the #1 reason for outages

- Find the right reliability/velocity balance for your product.
- Minimize unnecessary risk from changes.
- Don't (only) test in production.
- Use GitOps.
- Don't deploy on "Fridays" (for varying values of $Friday).
- Production freezes don't solve the underlying problem.

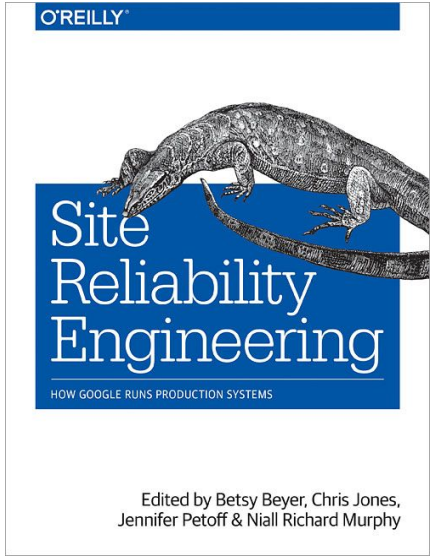
Outages are inevitable

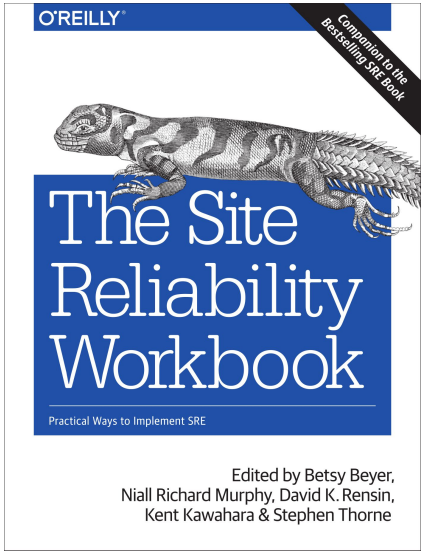
- You need change. Change has risks.
- The goal is not to prevent outages entirely, but to limit their overall cost.
- Try to mitigate first, root cause later.
- Be able to roll back your change quickly.
- Use written communication for incident management.
- Organizational transparency helps with root causing. Read code, not docs.

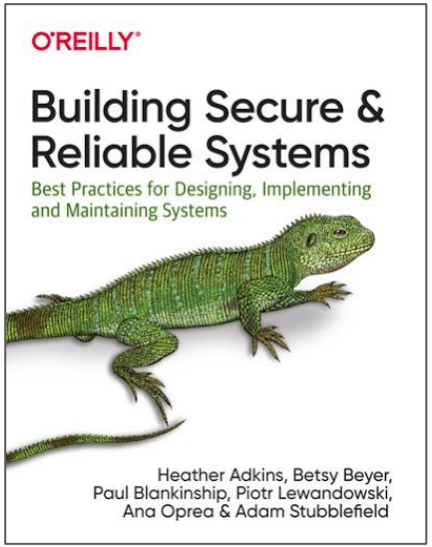
No Haunted Graveyards

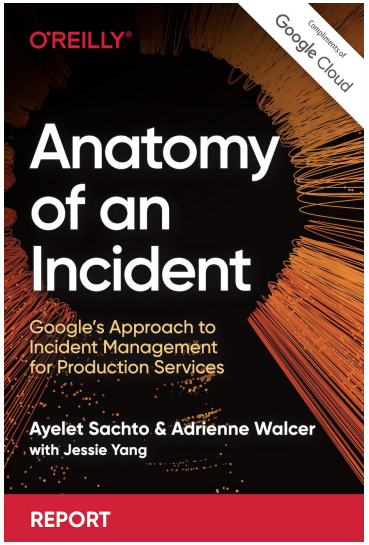
- A system can become so fragile and complex that no one dares to touch it.
- Complexity is a booby trap for change - complex systems need constant fixing.
- Don't accept neglect ("broken windows"). It's a slippery slope.
- Anyone can build complex systems - try building simple systems!

Follow us on Twitter: @googlesre. Find Google SRE publications—including the SRE Books, articles, trainings, and more—for free at sre.google/resources.









Book covers copyright O'Reilly Media. Used with permission.

Thank you!