Maturing Incident Management

Steve McGhee
smcghee@google.com

$DATE
Prepared for: $CUSTOMER

new speaker, who dis

- 10+ years operating distributed systems within Google (Android, Cloud, YouTube)
- SRE Lead - A deep understanding of SRE principles
- N years using Kubernetes and GKE (and Borg)
- Recent first-hand experience as a Cloud customer, migrating a complex, hybrid $XB business
  - including M&A properties, 10+ year legacy systems

Fire Fighting

Teams start out as **"volunteers"**

- Doing their best
- Works pretty well

Eventually, you *may* need a **professional, dedicated, funded(!) fire department.**

➔ **What** does this mean?
➔ **When** does the cost/value make sense?
➔ **How** to begin?

see also: https://noidea.dog/fires

What NOT to measure?

- incident count
  - declaring incidents **must be zero cost**
  - avoid the chilling effect or any *hesitation* in fast-moving env
  - any perceived social retribution for declaring "too early/often"

- MTTR, MTTF ← appealing, but misleading
  - "*if you doubled the incident count while the incidents follow roughly the same distribution, your system's reliability has clearly worsened, but your metric has not changed a lot.*"
  - https://sre.google/resources/practices-and-processes/incident-metrics-in-sre/

What to Measure

What **should** we measure instead?

Direct measures of reliability:

- SLO performance over week/month/quarter across services, teams
- "Learnings" about your system, reliability backlog generation
- Support case volume – "do customers really notice?"
- Revenue (eg: $/minute)
  - does outage loss get recovered immediately after?
- Speed/Agility: deployment time, etc (see: fourkeys)

Two Outage Types

Consider two types of outages: **Normal** vs **Huge**

For the Huge:

- "all hands on deck" / P0 / SEV1 / Code Red
- these can be **expensive** (time * people, distraction)

Balance freedom to escalate with **cost** by performing a process postmortem:

- How well did the escalation and incident response work?
- Ignore the actual incident details, separate postmortem

Adjust escalation norms as needed

Escalation: page < incident < INCIDENT

Escalating from a single-person page to an incident **brings benefits**:

1. **bring in help** - more eyes, more hands
2. **raise awareness** - others may be affected, may need to know (stakeholders, customers, adjacent services, dependencies)
3. **speed overall resolution** - accidental dependencies or side effects might spill onto unaware teams
4. makes an incident **discoverable, referenceable, reviewable**

Qualify response based on **Impact**, eg:
https://response.pagerduty.com/before/severity_levels/

So, What Makes an Incident?

"*Not every page, surely?!*" (but what if? consider SLO-alerts)

It's better to **declare early and close quickly**, than to fight an outage in the dark, then later apply the incident framework to a spreading incident.

A quick litmus test: Declare if **any** of the following apply:

1. Do you need to involve a **second team** in fixing the problem?
2. Is the outage **visible** to customers?
3. Is the issue unsolved even after **an hour**'s concentrated analysis?

Norms to establish early

Let on-calls know they are expected to **delegate and escalate** during an incident

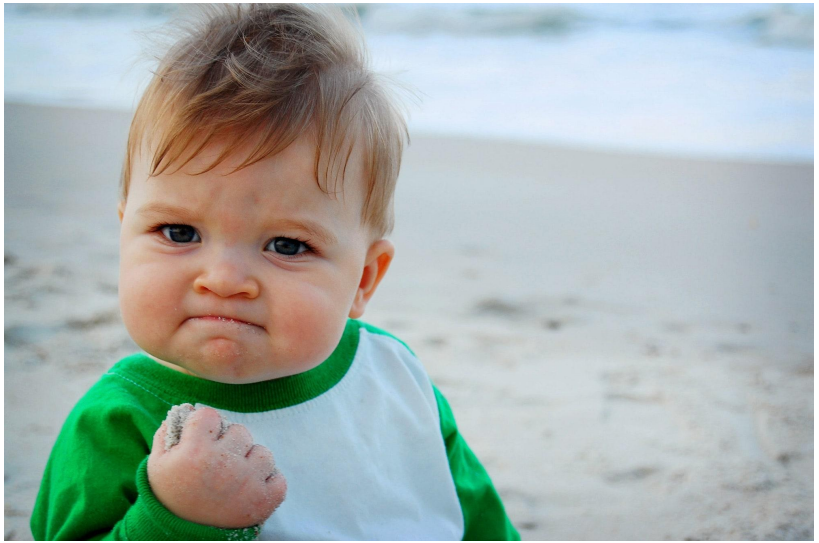
Encourage a **mitigation-first response**

Establish an expected **Command Post**

- ➔ generic: #panic, #teamname-panic
- ➔ specific: #inc12345

Practice **live collaborative documentation**

Practice **active handoffs** of state



Introduce the Incident Command Model

Developed by Fire Departments, Emergency Services.

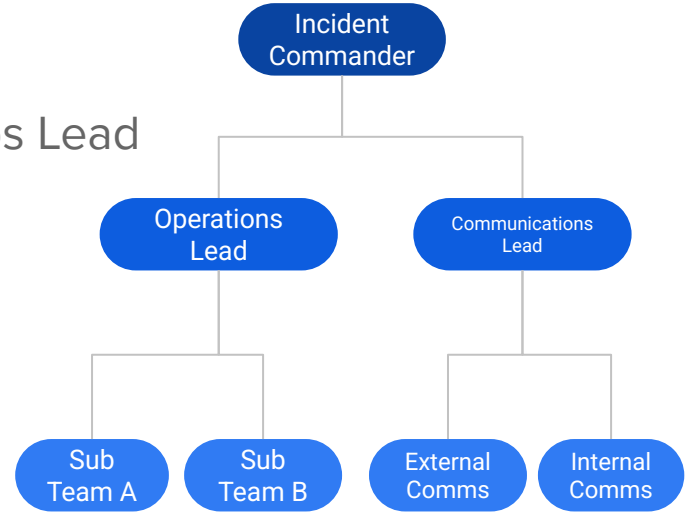
Adapted for SRE – Incident Management At Google (IMAG)

Coordinate / Communicate / Control

Primary roles: Incident Commander, Comms Lead, Ops Lead

https://sre.google/workbook/incident-response/

https://response.pagerduty.com/



Introduce Training

1. **Intro to ICM**
   - This is for everyone: oncall, not-oncall, managers, PMs, everyone.
   - Why are we doing this, define roles/terms.
2. **Tooling for ICM**
   - For oncall team members
   - Escalation process, intro to consistent naming schemes, backup methods
   - Sample live tracking documents, communication methods
   - Homework: go run a mock incident, today! (use a "dev" namespace in tools)
3. **How to be an IC.**
   - A subset of oncall, folks with experience, expectation to "keep cool"
   - Not limited to Managers! Must be hands-on.
   - Never an Executive.

Be Prepared!

You're not ready until you prepare

- Make team contact lists - paper "panic card" for larger orgs and teams
- Decide on escalation paths: when to escalate and how.
  - It won't be perfect, that's fine. Write it down.

Drill Drill Drill

- **DiRT** - *Yearly/Quarterly* – Org-wide synchronous week of Disaster Test
  - `[DiRT test] Aliens Have Landed, Need capacity NOW!!`
- **WoM** - *Weekly* – oncall team specific. spin the wheel, practice on previous outages, or make one up! (think: Dungeons & Dragons)
  - great for new-to-oncall to get up to speed, improve tooling, slow things down

Tip: give historical incidents memorable code names

Big Incidents

"whoa" "oh no"

Escalate early.

Establish an "A Team"

- coordination, not domain expertise
- broad access power (root)
- purchase authority
- empowerment to tell other teams to engage, what to do
- may ask to take over IC, will always accept of you offer.

https://www.usenix.org/conference/lisa21/presentation/walcer
https://response.pagerduty.com/before/complex_incidents/

What is Big?

in an incident, when to further **escalate**?

- significant user / business **impact**
- likely to involve **multiple teams**
- has **potential to expand**, get worse
- has been going for **30+ minutes** already
- **"feels bad"** but you're not sure why

Longer Incidents

- shared documents (not just the postmortem)
  - mitigation plans
  - purchasing budgets, approvals
  - duty coverage spreadsheets
- **explicit handovers**
  - **planned** handovers! (work/sleep schedules)
- pre-schedule post-incident reviews of process, notes, postmortem
- be able to walk away at agreed all-clear condition
  - the IC declares an all-clear

A Note on the Comms Lead:

Not everyone can/should be Comms Lead.
(Especially external comms)

Duties include:

- statuspage updates
- watches the clock for providing updates
- proactive notification of customers
- briefing Execs
- working with Press/Media

Need to work well with Support, Sales, Customers, Execs, Marketing

What to measure – your homework!

2 weeks: measurement, identify **cases where this would have helped**

3 months: plan and hold a **team workshop** - propose model, get feedback

6 months: start training, measure the "right" metrics over time, see value

other:

book club: sre workbook, secure/reliable – read/discuss a chapter every 2 weeks

Quick Review

- "Volunteers" work up to a point, time to formalize
- Measure reliability, not response (don't count incidents!)
- Train everyone on basics, Train specialists on specifics
- Drill ! (WoM weekly, DiRT quarterly)
- Introduce an A-Team for LARGE incidents
- Manage Big/Long incidents differently
- Review outages regularly and deeply

*"if you have a **better** and **richer understanding** of the incident, you will have more productive things to do about it in the future."* - John Allspaw
https://croz.net/news/0800-john-allspaw-on-incidents-teams-and-learning-organizations/

Investing in Fire Trucks

You're investing in a Fire Truck, train your team to be Professionals.

Thanks for your time

This is a big step!

We're here to help.