Google

Reporting on Reliability

Improving stakeholder conversations

sre.google • twitter.com/googlesre

<Presenter Name>
<Presenter Role>
@<Twitter Handle>

Site Reliability Engineering

principles

Incidents happen.

The goal is not 100% uptime.

Reliability matters.

Reliability takes work.

Reporting facilitates decision making

Decisions may include...

- Resource allocation
  - Human resources
  - Capital investments
- Prioritization
  - Feature A vs. Feature B vs Reliability Project A vs. Reliability Project B
- Communications
  - External
  - Internal

What the heck?!?

What the heck?!?

What was that??

What the heck?!?

What was that??

How’s it going?

What the heck?!? ← Incident status

What was that??

How's it going?

What the heck?!? ← Incident status

What was that?? ← Postmortem

How's it going?

What the heck?!? ← Incident status

What was that?? ← Postmortem

How's it going? ← Periodic reviews

What the heck?!? ← Incident status

What was that?? ← Postmortem

 How's it going? ← Periodic reviews

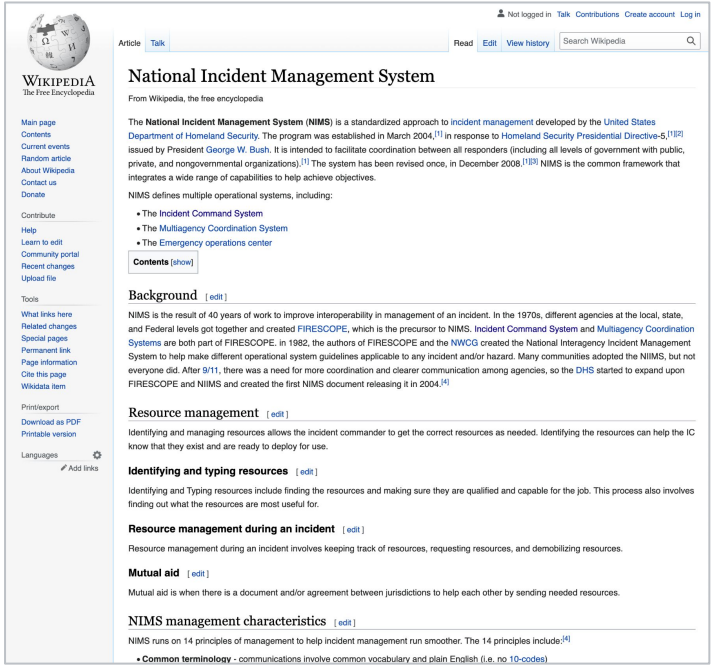
What the heck?!?

Real-time incident status

What the heck?!? Real-time incident management

Assign clear, specific roles

- Incident commander
- Communications manager
  - Internal
  - External



Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Languages
Add links

National Incident Management System

From Wikipedia, the free encyclopedia

The **National Incident Management System** (**NIMS**) is a standardized approach to incident management developed by the United States Department of Homeland Security. The program was established in March 2004,[1] in response to Homeland Security Presidential Directive-5,[1][2] issued by President George W. Bush. It is intended to facilitate coordination between all responders (including all levels of government with public, private, and nongovernmental organizations).[1] The system has been revised once, in December 2008.[1][3] NIMS is the common framework that integrates a wide range of capabilities to help achieve objectives.

NIMS defines multiple operational systems, including:

- The Incident Command System
- The Multiagency Coordination System
- The Emergency operations center

Contents [show]

Background [edit]

NIMS is the result of 40 years of work to improve interoperability in management of an incident. In the 1970s, different agencies at the local, state, and Federal levels got together and created FIRESCOPE, which is the precursor to NIMS. Incident Command System and Multiagency Coordination Systems are both part of FIRESCOPE. in 1982, the authors of FIRESCOPE and the NWCG created the National Interagency Incident Management System to help make different operational system guidelines applicable to any incident and/or hazard. Many communities adopted the NIIMS, but not everyone did. After 9/11, there was a need for more coordination and clearer communication among agencies, so the DHS started to expand upon FIRESCOPE and NIIMS and created the first NIMS document releasing it in 2004.[4]

Resource management [edit]

Identifying and managing resources allows the incident commander to get the correct resources as needed. Identifying the resources can help the IC know that they exist and are ready to deploy for use.

Identifying and typing resources [edit]

Identifying and Typing resources include finding the resources and making sure they are qualified and capable for the job. This process also involves finding out what the resources are most useful for.

Resource management during an incident [edit]

Resource management during an incident involves keeping track of resources, requesting resources, and demobilizing resources.

Mutual aid [edit]

Mutual aid is when there is a document and/or agreement between jurisdictions to help each other by sending needed resources.

NIMS management characteristics [edit]

NIMS runs on 14 principles of management to help incident management run smoother. The 14 principles include:[4]

- **Common terminology** - communications involve common vocabulary and plain English (i.e. no 10-codes)

What the heck?!? Real-time incident management

Use appropriate communication channels

- Avoid conference calls
  - Noisy and lossy
  - Single threaded
- Chat/IM
  - Multi threaded
  - Allows multicast as well as small groups
  - History can be preserved
- Live docs and collaborative systems
  - Capture (but don't publish) in real time

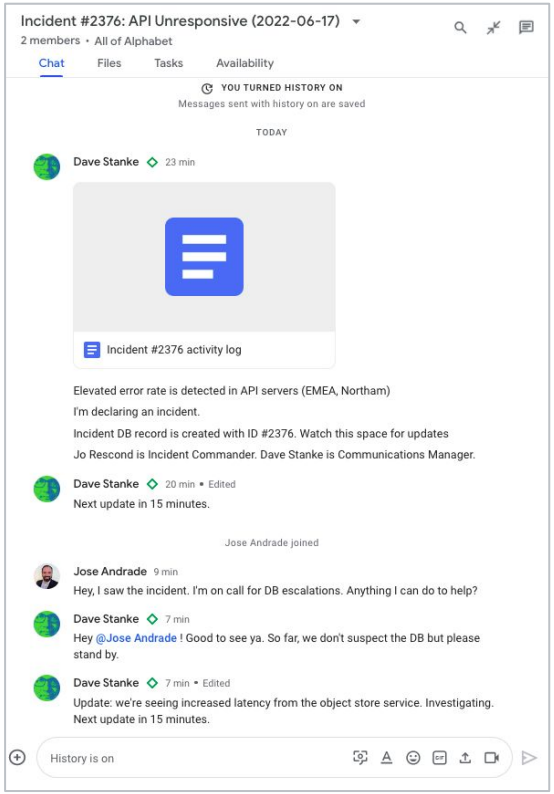
What the heck?!? Real-time incident management

Be predictable

- Regular updates with new information
- Each update includes “next update in {x} minutes”

Most people should mostly listen

- Core responders broadcast to stakeholders
- Others can offer to help
  - But wait until acknowledged to take action



What was that??

Postmortem

What was that?? Postmortem

- Capture data and experiences
- Resolve ambiguities
- Complete the narrative
  - For consumption by others
- Plan next steps

> **Incidents are unplanned investments**, and they are also opportunities. Your challenge is to maximize the ROI on the sunk cost. To do that, the organization has to invest in really exploring and understanding these events, and share that understanding broadly and over time.*
>
> **—John Allspaw**

* John Allspaw, DevOps Enterprise Summit 2018, based on stella.report

What was that?? Postmortem

Blamelessness is key

- Criticize systems, not people
- Assume that everyone involved in an incident had good intentions
- “Human” errors are systems problems

  “Jamie shouldn't have done that” ⟶
  “The system *shouldn't have allowed* an engineer to do that”

- A culture of blame leads to hiding of problems and poor morale

What was that?? Postmortem

Documenting the incident

- Participants are cross-functional stakeholders
  - *The ones who were involved in the incident!*
- Provide time and space for open communication
- Be Comprehensive
  - Status (e.g. “Complete, action items in progress”)
  - Summary
  - Impact
  - Contributing factors
  - Trigger
  - Resolution
  - Action items
  - Timeline

What was that?? Postmortem

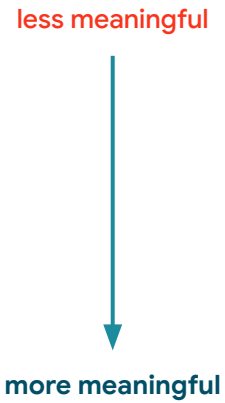
Action items (TODOs)

- There should probably be several
  - Is there *really* only one problem to fix?
  - Look beyond specific bugs, to find systemic contributing factors
- Seek opportunities to improve...
  - ...detection
  - ...mitigation
  - ...prevention
- Assign them, prioritize them, track them
  - The incident isn't really over until all the AIs are completed

What was that?? Postmortem

Assessing impact: How bad was it?

- It happened
- It lasted for {x} minutes
- It affected systems A, B, C
- It impacted users attempting to {do_thing}
- It caused $X in lost revenue



What was that?? Postmortem

Independent review

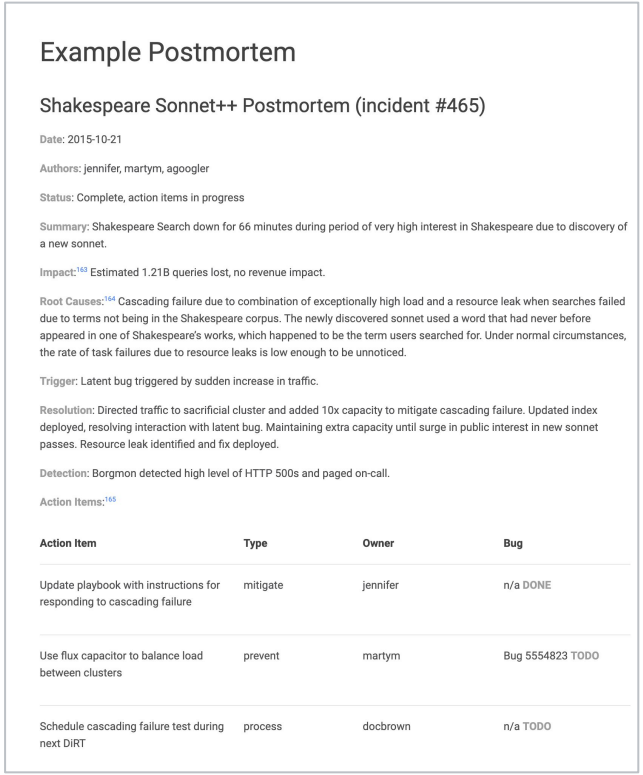
Before publishing a postmortem report to stakeholders, solicit peer review to ensure it's clear and comprehensive.

Make it easier by hosting regular office hours.

What was that?? Postmortem

Publishing and distribution

- Executive Summary
  - TL;DR
- Push to known channels
- Publish to known locations
- Schedule a readout
- Invite questions



Example Postmortem

Shakespeare Sonnet++ Postmortem (incident #465)

Date: 2015-10-21

Authors: jennifer, martym, agoogler

Status: Complete, action items in progress

Summary: Shakespeare Search down for 66 minutes during period of very high interest in Shakespeare due to discovery of a new sonnet.

Impact:[163] Estimated 1.21B queries lost, no revenue impact.

Root Causes:[164] Cascading failure due to combination of exceptionally high load and a resource leak when searches failed due to terms not being in the Shakespeare corpus. The newly discovered sonnet used a word that had never before appeared in one of Shakespeare's works, which happened to be the term users searched for. Under normal circumstances, the rate of task failures due to resource leaks is low enough to be unnoticed.

Trigger: Latent bug triggered by sudden increase in traffic.

Resolution: Directed traffic to sacrificial cluster and added 10x capacity to mitigate cascading failure. Updated index deployed, resolving interaction with latent bug. Maintaining extra capacity until surge in public interest in new sonnet passes. Resource leak identified and fix deployed.

Detection: Borgmon detected high level of HTTP 500s and paged on-call.

Action Items:[165]

Action Item	Type	Owner	Bug
Update playbook with instructions for responding to cascading failure	mitigate	jennifer	n/a DONE
Use flux capacitor to balance load between clusters	prevent	martym	Bug 5554823 TODO
Schedule cascading failure test during next DiRT	process	docbrown	n/a TODO

sre.google/sre-book/example-postmortem

How’s it going?

Periodic reviews

How's it going? Periodic reliability reviews

- Where decisions are made!
  - Learn from the past
  - Plan for the future
- Scheduled and templated
  - Twice a year
    (or thereabouts; adjust as needed)
- Attendees:
  - Decision makers (management; executives)
  - Technical leads from the team
    - Not just SRE; all "owners," including Devs, Product, etc.
  - SRE experts (from beyond the team)

How's it going? Periodic reliability reviews

Look back

Look ahead

Look all around

How's it going? Periodic reliability reviews

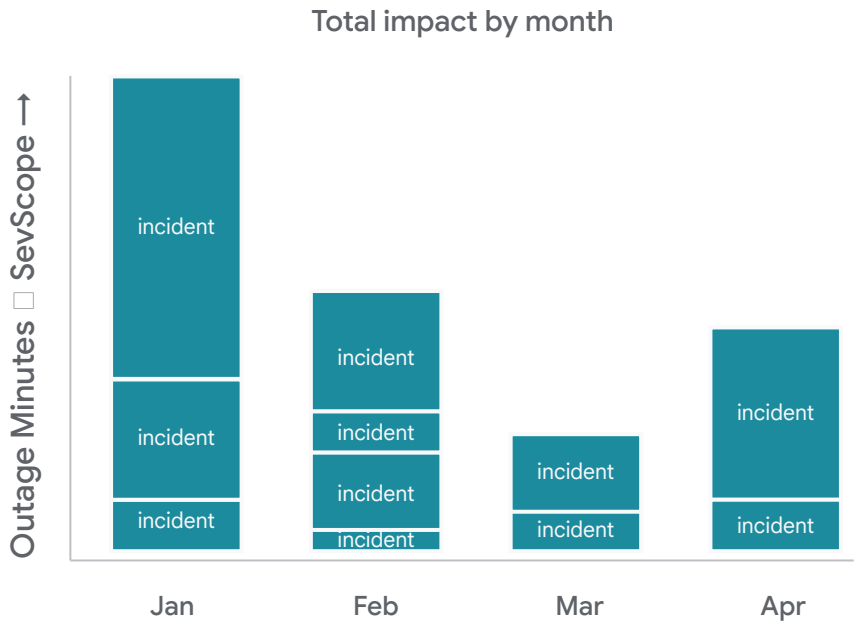
Look back: SLO rollup

- Restate SLO targets
  - Assess SLO compliance: targets hit/missed
- Analyze upstream and downstream dependencies
  - This service's impact to others; other services' impact to this
- Consider revising targets

How's it going? Periodic reliability reviews

Look back: Incident report

- Don't focus on incident counts
  - Each incident is unique
- Do focus on the aggregated impact of incidents
  - Revenue lost
  - Error budget burned
  - Blast radius
  - Teams/services affected
  - *Look for patterns*



How's it going? Periodic reliability reviews

Look back: Incident report

- Major incidents
  - Where did they happen?
    - Are there trends? Clusters?
    - Specific teams that struggle?
  - What was the impact?
    - Is it concentrated on particular regions? Customers? User types?
  - Action Items
    - Are they getting done?

Beware of recency bias

How's it going? Periodic reliability reviews

Look back: Team health

- **People** Does the team have the right mix of skills? What training can we offer?
- **Cognitive Load** Is the team at capacity, or can they onboard additional services?
- **Toil** Are we doing the right amount? Are we learning from it?
- **Interrupts** Consider a paging "budget" (e.g. max 2 per on-call shift)
- **Morale** A happy team is a reliable team

How's it going? Periodic reliability reviews

Look ahead

- Forecast: demand
  - What capacity will be needed?
  - What other factors will become relevant?
- Forecast: work to be done
  - What's in the reliability backlog? What might prevent it from getting prioritized?
- Plans
  - Are stakeholders expecting improved reliability? Can we achieve it?
  - What launches are coming, for this team or related teams?
  - What known future bottlenecks/deprecations/etc. can we prepare for?
  - How does the team's work align to (ongoing or emerging) organizational strategy?

How's it going? Periodic reliability reviews

Look all around

- Dependencies
  - Who is dependent on this team? Are we supporting them well?
  - Who does this team depend on? Are they supporting us well?
- Platforms and tools
  - Does this team use common standards?
  - Does this team contribute back to the ecosystem?
    - PRs to upstream projects; creating/maintaining tools; tech talks or consultation
- Learning
  - Who can learn from this team? Who can this team learn from?

How's it going? Periodic reliability reviews

Requests and proposals → decisions!

- **Do we need to change our plans?**
  - Product Roadmap
  - Communications
  - Protocols
  - Resourcing
  - Team structure
  - Engagement model
- **Meeting outcome: updated plans**
  - e.g. OKRs or other planning artifacts; staffing; budgets

Reliability is hard.

When we talk about reliability, it helps to...

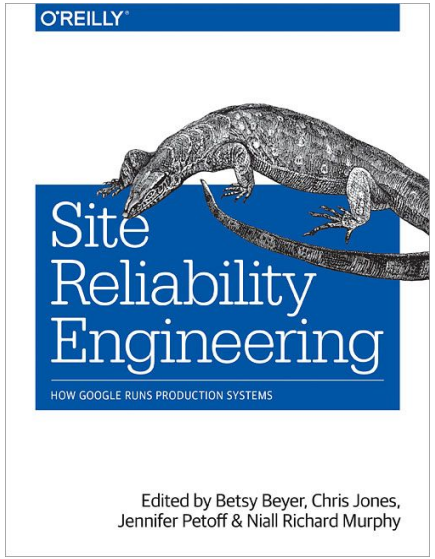
- Respect the inevitable
  - Incidents will happen
- Practice blamelessness
  - Assume good intent
- Share insights broadly
  - Nurture a learning community
- Reflect and iterate

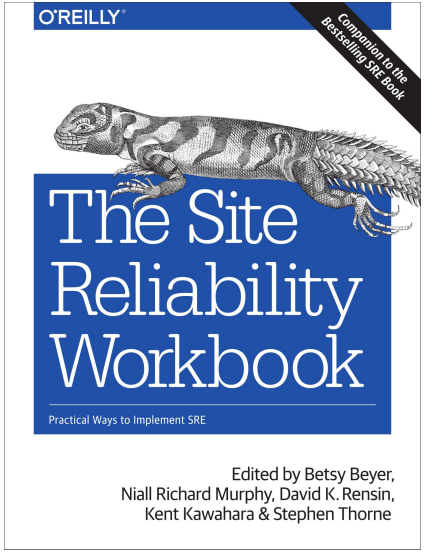
A “near miss:” silver lining?

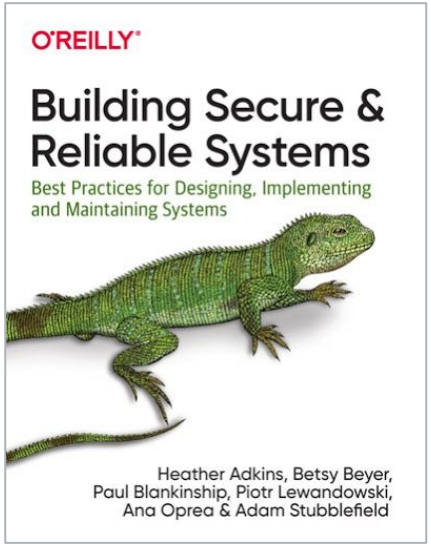
A “near miss:” solid gold.

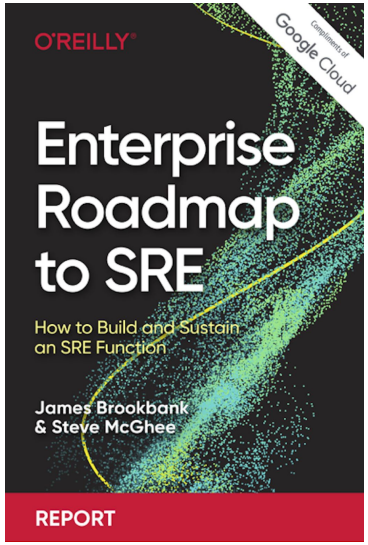
May all your incidents be interesting.

Find Google SRE publications—including the SRE Books, articles, trainings, and more—for free at sre.google/resources.









Emergency Incident Response

Planet-Scale Distributed Systems

Service Level Objectives (SLOs)

Systems Engineering

Global Storage

Load Balancing

Monitoring

Availability

Embracing Risk

Blameless Failures

Site Reliability Engineering

Software Engineering

Automation

"Hope Is Not A Strategy"

sre.google

Extra bits

Calculus, anyone?

- Individual errors
  - ...are the domain of machines
- Incidents (1st derivative of errors)
  - ...are the domain of SREs
- Trends (2nd derivative of errors)
  - ...are the domain of leadership