Google

Organizational Design
for
Technical Emergency Response
in Distributed Computing Systems

- A. Walcer, Google, Inc.
- A. Perry, Google, Inc.

Site Reliability Engineering

Abstract

When a company critically relies on the ongoing functioning of a complex and highly interconnected technical stack, support of that stack implies that appropriate personnel be reliably available to troubleshoot and correct issues that occur. These personnel will be referred to as responders. When the scope of a technical stack grows beyond one person's capacity to understand and maintain state, we split up the technical stack such that multiple responders can each provide coverage on a single component of the whole stack. Such a highly interconnected system-of-systems (SoS) allows production issues to cascade throughout wide swaths of the SoS, or sneak in between system-to-system (StS) boundaries. We will here explore one private industry implementation of a responder group designed to respond to emergent distributed computing SoS failures. In contrasting the functions of component responders and SoS responders, we demonstrate that the component ownership skillset is distinguishable from the core skill set of an SoS responder. Technical organizations can benefit from setting up SoS response to enable expedient distributed system outage mitigation.

Google's SRE Disaster Team

Prediction

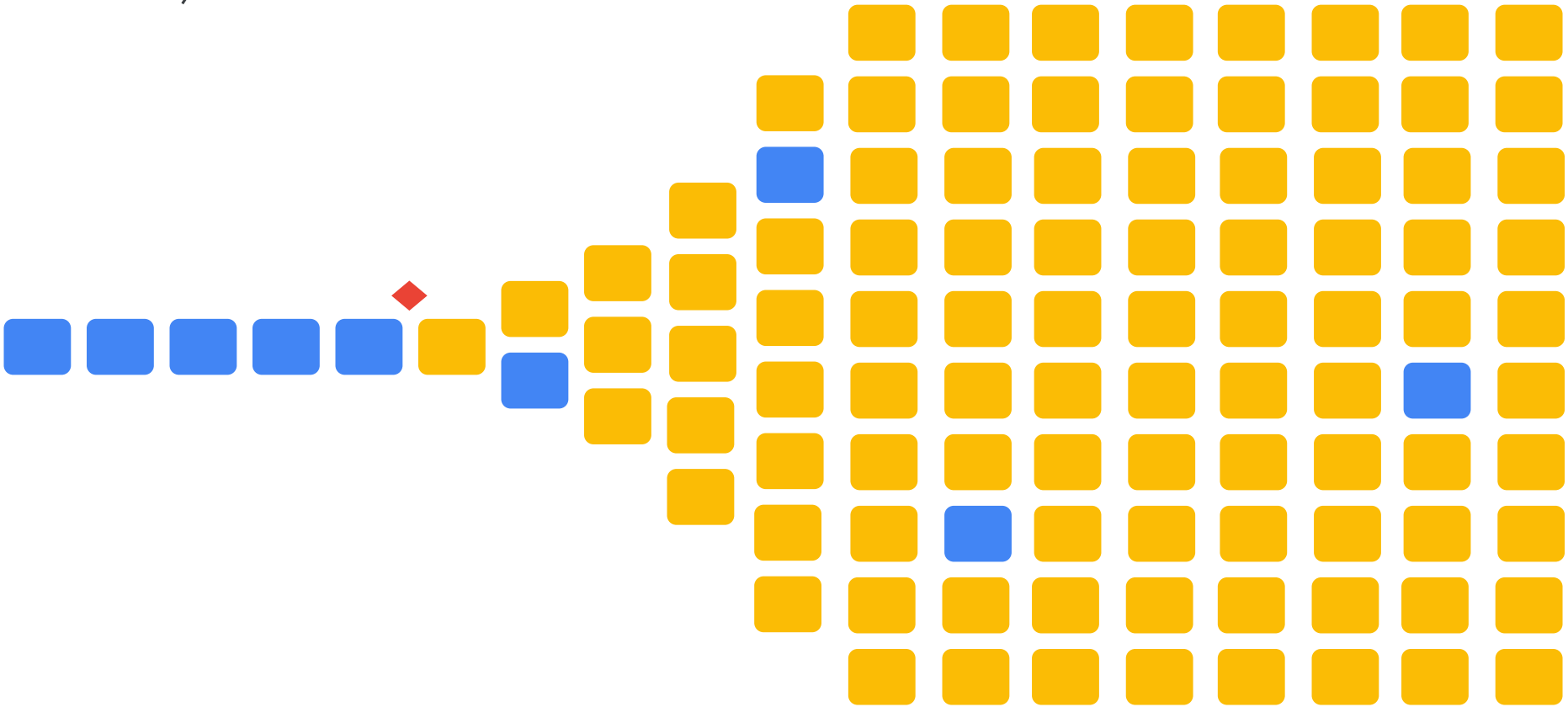
Prevention

Preparedness

Response

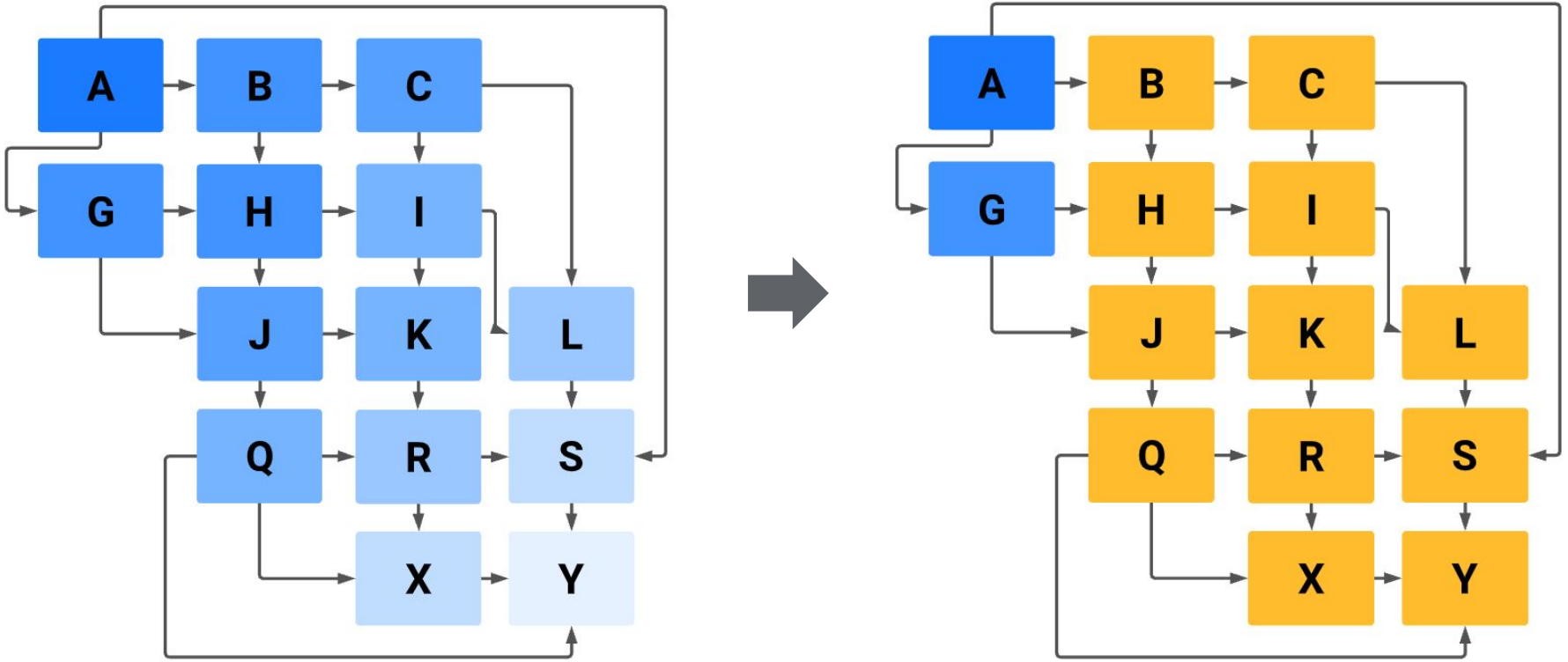
The Mayan Apocalypse

June 2, 2019



June 2, 2019

June 2, 2019



June 2, 2019

June 2, 2019


The Washington Post
Democracy Dies in Darkness
Tech
Consumer Tech
Future of Transportation
Innovations
Internet Culture
Space
Technology
Google Calendar is back online after 3-hour outage
CBS NEWS
Google Cloud outage hits YouTube, Gmail and more
BY APRIL SIESE
UPDATED ON: JUNE 2, 2019 / 8:22 PM / CBS NEWS
THE VERGE
TECH
SCIENCE
GOOGLE
TECH
SNAPCHAT
30
Google recovers from outage that took down YouTube, Gmail, and Snapchat
Google Cloud problems hit a number of se
By Tom Warren | @tomwarren | Jun 2, 2019, 4:02pm EDT
The New York Times
Subscribe now for $1 | Log In
Google Disruptions Affect Gmail, YouTube and Other Sites

Responder, yes,
but which Role of Responder?

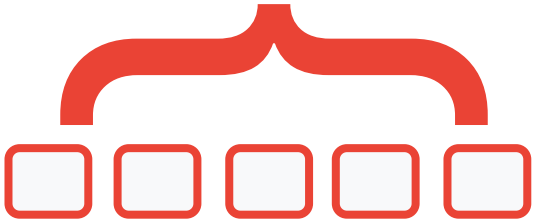
Incident responders in SRE

Component Responders

Incident Responders oncall for one component or system within Google's overall technical infrastructure.

System-of-Systems (SoS) Responders

Incident Responders oncall to support incidents that span multiple component systems, incidents that fall between system boundaries, or anything that gets messy.



Component Response

Component responders

Single-system experts....

- Well-versed in the problem space
- Expert troubleshooters
- Practiced in implementing mitigation strategies during a crisis
- Continual access to the tools/systems required to perform emergency response
- Handle stress well and think clearly during a crisis

Component responders

At Google, these are divided between:

Infrastructure components

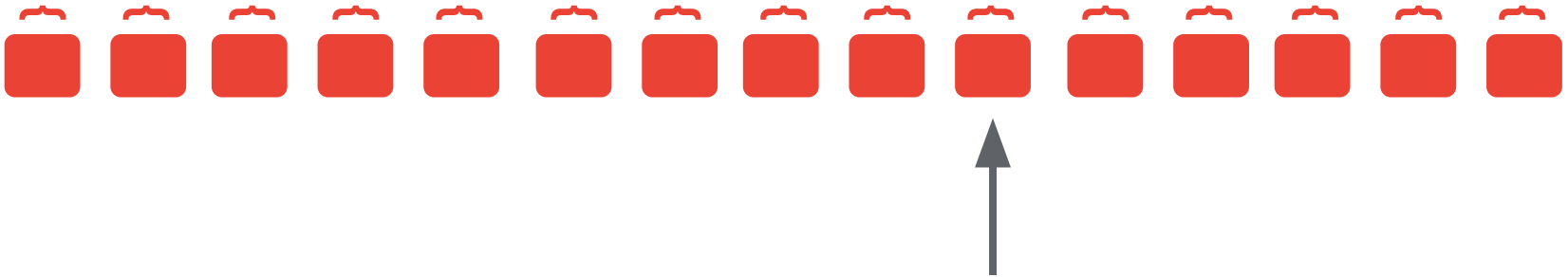
Product Service components

Internal Service components

Component Responders

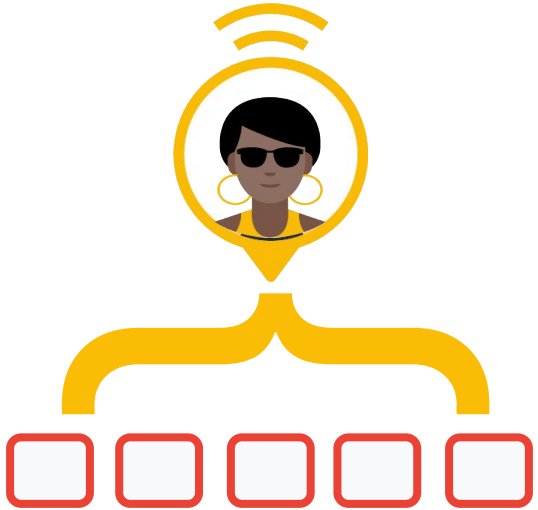
Systems-of-Systems Response

From components to bigger picture



If this **component** is triggering an issue **across other components,** how do we coordinate between them?

System-of-systems responders



Multi-system incident managers....

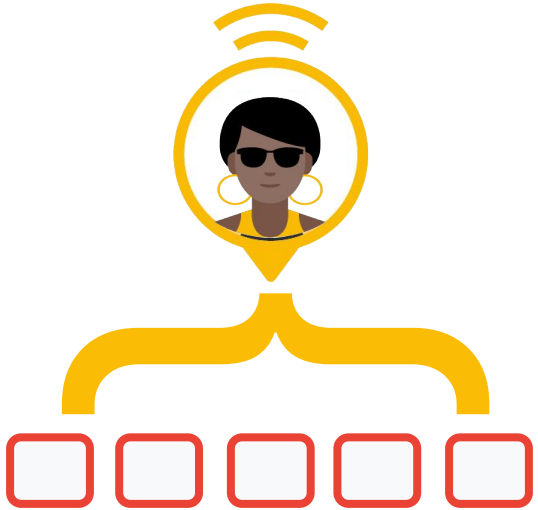
- Skilled generalists
- Holistically focused
- Organize others
- Command complex situations
- Diagnose systemic behaviors
  - and identify root issues
- Focused on scaling response
  - and communicating widely

System-of-systems responders

At Google, these are divided between

The **Product-Focused incident response teams** (IRTs) that...
take responsibility for incidents that are pervasive across broad swaths of a specific product, or similar products

and ...



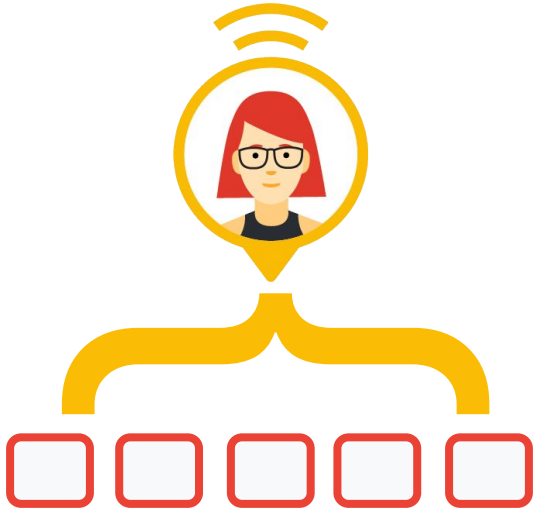
System-of-systems responders

At Google, these are divided between

The **Product-Focused incident response teams** (IRTs) that...
take responsibility for incidents that are pervasive across broad swaths of a specific product, or similar products

and

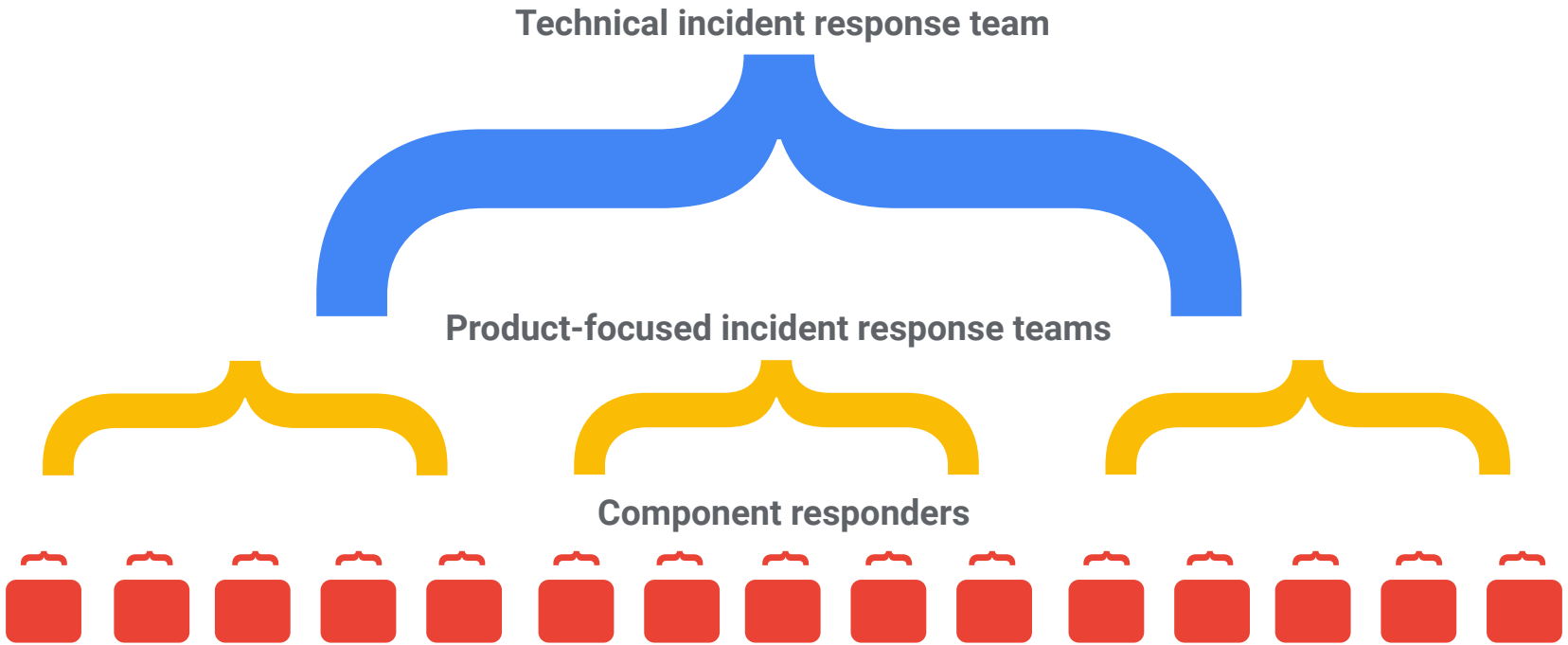
The **Technical incident response team** (IRT) that...
responds to and helps coordinate, mitigate and/or resolve major service outages across Google (often due to incidents rooted in shared infrastructure)



System-of-systems responders

EVERYTHING IS BROKEN AND NOTHING IS WORKING ⚡

System-of-systems responders



Enabling Factors

Common Protocol

All responders use the same incident management protocol, allowing for role clarity and shared rules of engagement

Enabling Factors

Common Protocol

All responders use the same incident management protocol, allowing for role clarity and shared rules of engagement

Trust

Responders are given the authority to handle the incident, without seeking approvals for every action

Enabling Factors

Common Protocol

All responders use the same incident management protocol, allowing for role clarity and shared rules of engagement

Trust

Responders are given the authority to handle the incident, without seeking approvals for every action

Respect

Creating and maintaining psychological safety is everyone's responsibility

Enabling Factors

Common Protocol

All responders use the same incident management protocol, allowing for role clarity and shared rules of engagement

Trust

Responders are given the authority to handle the incident, without seeking approvals for every action

Respect

Creating and maintaining psychological safety is everyone's responsibility

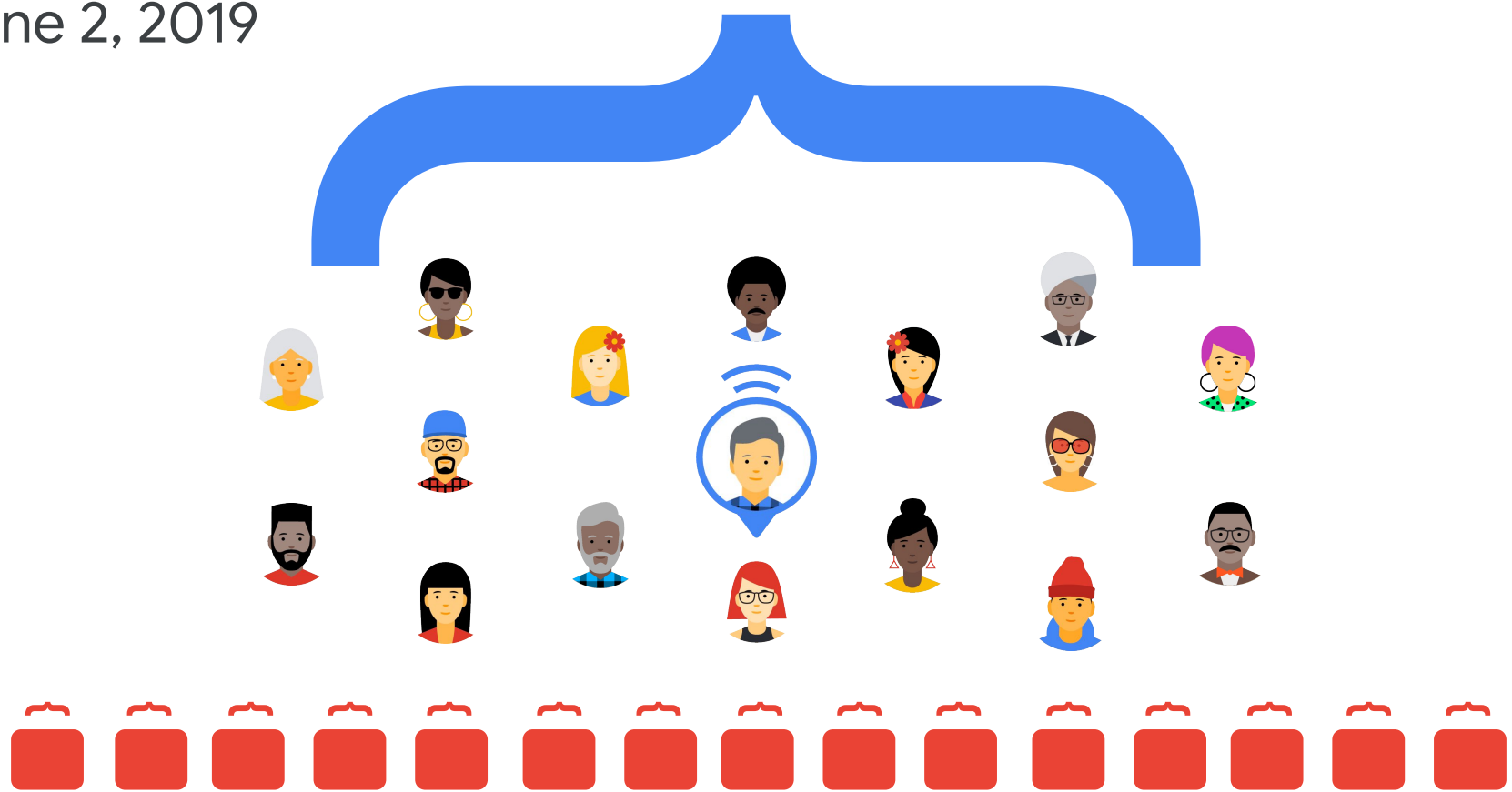
Transparency

Incidents are reported openly across the company

Technical Outage, Incident Response

Back to June 2, 2019

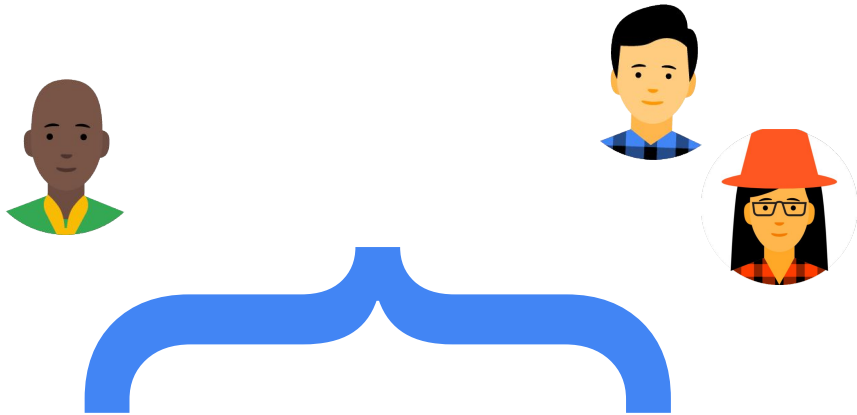
June 2, 2019



June 2, 2019

Tech IRT members....

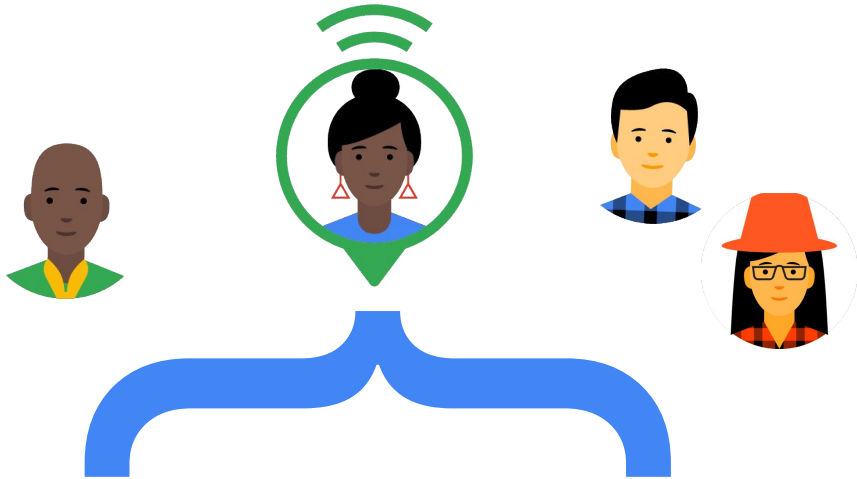
- Formally assume incident command
- Assess the current state of an incident
- Organize people to coordinate the moving parts of the response
- Set priorities and delegate tasks
- Secure additional resources where needed
- Remove administrative and communications burdens from the folks that can implement mitigations



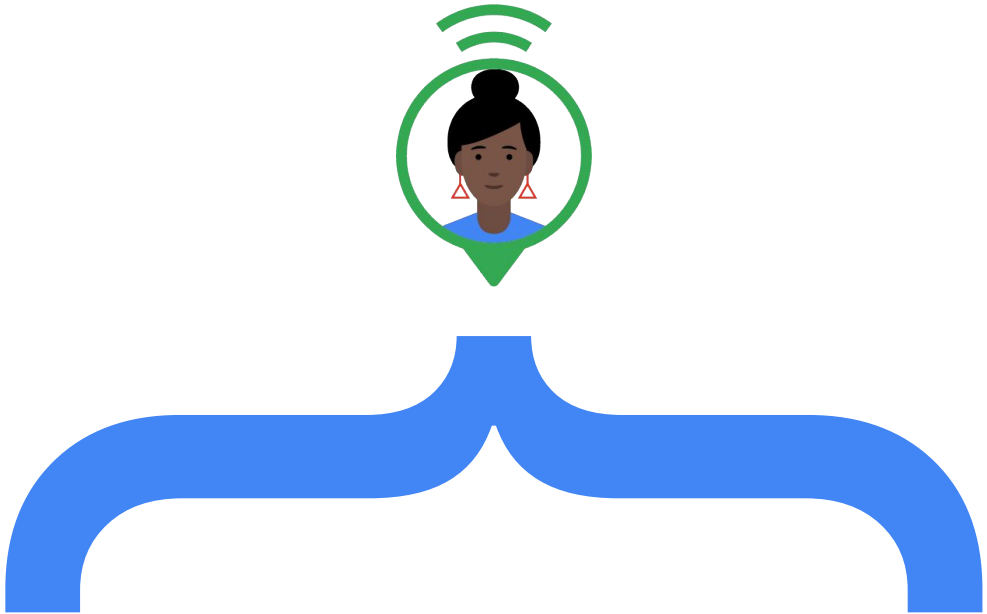
June 2, 2019

Tech IRT members....

- Formally assume incident command
- Assess the current state of an incident
- Organize people to coordinate the moving parts of the response
- Set priorities and delegate tasks
- Secure additional resources where needed
- Remove administrative and communications burdens from the folks that can implement mitigations



June 2, 2019



The Aftermath: June 2, 2019

- Incredibly detailed postmortem
- Spin-off engineering work to address the root cause and trigger conditions (to prevent re-incident)
- Rewarding the people involved

Thank you, any questions?

With Thanks To...

Patrick Bernier
Kieran Broadfoot
Alex Matey
John Truscott Reese
Benjamin Treynor-Sloss
Murali Suriar
Aaron Topal
Todd Underwood
Andrew Widdowson

Citations

[1] “Google Data Center FAQ.” Data Center Knowledge, 19 Mar. 2017, https://www.datacenterknowledge.com/data-center-faqs/google-data-center-faq-part-3.

[2] Aleksandra. “63 Fascinating Google Search Statistics.” SEOtribunal, 26 Sept. 2018, https://seotribunal.com/blog/google-stats-and-facts/.

[3] “Incident Command System Resources.” FEMA, The U.S. Department of Homeland Security, 26 June 2018, https://www.fema.gov/incident-command-system-resources.

[4] Beyer, Betsy. Site Reliability Engineering: How Google Runs Production Systems. OReilly Media, 2016.

[5]“How Google Protects Your Data: Data Access and Restrictions.” Google Cloud Security and Compliance, Google Cloud, https://gsuite.google.com/learn-more/security/security-whitepaper/page-7.html.

[6] Treynor Sloss, Benjamin. “An Update on Sunday's Service Disruption.” Inside Google Cloud, Google Cloud, 3 June 2019, https://cloud.google.com/blog/topics/inside-google-cloud/an-update-on-sundays-service-disruption.