Who to blame for all of your problems

Conducting blameless postmortems

About Me

Ben Berry

DevOps Engineer/Site Reliability Engineer

@bengerman13

me@bengerman.com

What is a postmortem?

What is a postmortem?

It’s actually two things:

- A document containing the details about an incident
- The meeting held to create such a document

Why should we conduct postmortems?

- Complex systems fail in complex ways
- Engineers have an incomplete and inaccurate view of complex systems

Postmortems let you listen to your system

What is Blamelessness?

Assume every person made the correct decision at every point along the way, given the information that was available to them at the time.

Why blameless postmortems?

- Honesty and Transparency
- Failure is an outcome, not a behavior
- What makes sense to one person now will probably make sense to someone else in the future

Anonymity != Blamelessness

- “Trigger: An engineer powered down the main fooserver”
  - Which engineer?
  - *Everyone glares at James*
- “Trigger: The main fooserver was powered down”
  - How did that happen?
- “Trigger: James powered off the main fooserver”

Accountability Without Blame

Blameful

- Revoke access
- Require approvals
- Fire them

Blameless

- Keep experienced engineers close to the problem
- Spread knowledge and insight
- Apply knowledge gained from incidents

“But my work is too important!”

“Conditions shape decisions and actions, and revealing these conditions will help the agency design … a more robust and resilient system”

- US Forest Service

https://wildfirelessons.connectedcommunity.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=d2166d55-e69c-1d8c-a2bb-5c8b8a272a7f

National Transportation Safety Board

"Our sole purpose is to determine how and why this accident occurred and what can be done to prevent similar occurrences in the future."

https://www.ntsb.gov/investigations/process/Documents/MajorInvestigationsManual.pdf

Getting started

Assume every person made the correct decision at every point along the way, given the information that was available to them at the time.

Postmortem parameters

- Any time an outage is declared
- Any time service is degraded for a certain amount of time
- Any time a human needs to intervene to correct a process that should be automatic
- Any time a stakeholder requests it

General Summary

Date: 2017-07-30
Authors: benb
Status: Complete
Summary: All instances of the web server stopped responding to requests for approximately 1 hour
Impact: Users were unable to view some content on the Visual Guide for 1 hour
Detection: Internal user (Graham) tried to access the page but could not

Causes

Root Causes:

- Incompatible versions of docker and docker-compose caused stdout to stop accepting input after a certain number of bytes were sent to it
- Java process writes logs synchronously
- We never test this service using docker-compose until we release
- Alerts were misconfigured, so we did not know about individual instances becoming unhealthy

Trigger: More verbose logging and longer time between deployments meant we actually got enough logs written to trigger this bug

Fixing it

Resolution:

- Turned on instance termination for failed health checks
- Pinned docker-compose and docker versions in ansible deployment scripts

Action Items	
Description	**Jira Link**
Support docker-compose in staging	JIRA-1234
Fix and test health check alarms	JIRA-5678

What we learned

What went well:

- Changing health check behavior to termination quickly made the system consistently available for users

What went wrong:

- We don’t have a staging environment that matches production
- We never tested alerting

Where we got lucky:

- Internal users caught this before public launch
- Cloudfront caching hid most of the errors

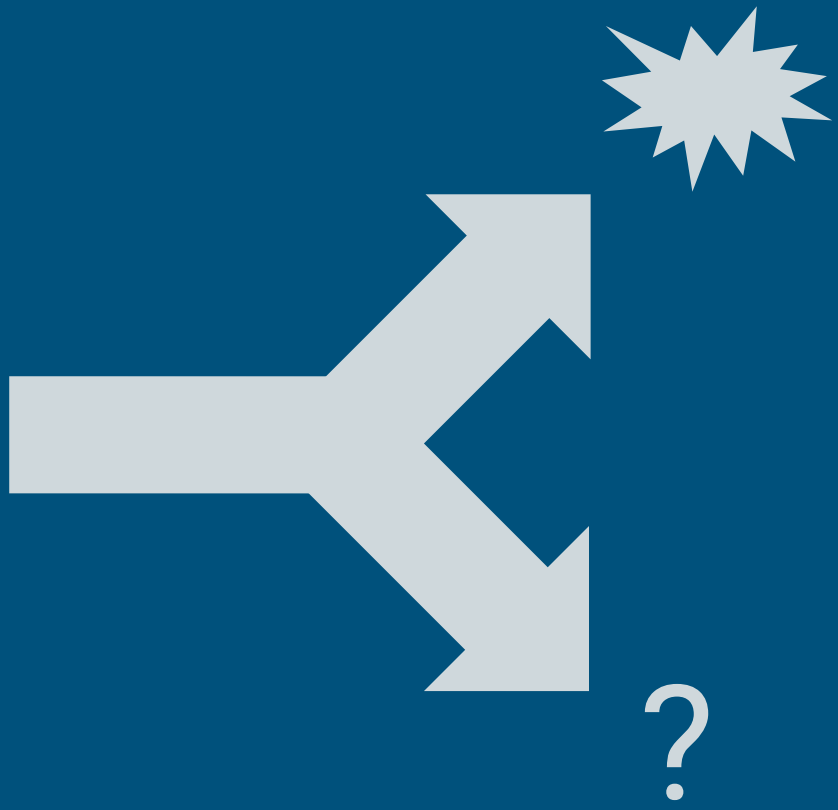
Before the meeting

Running the meeting

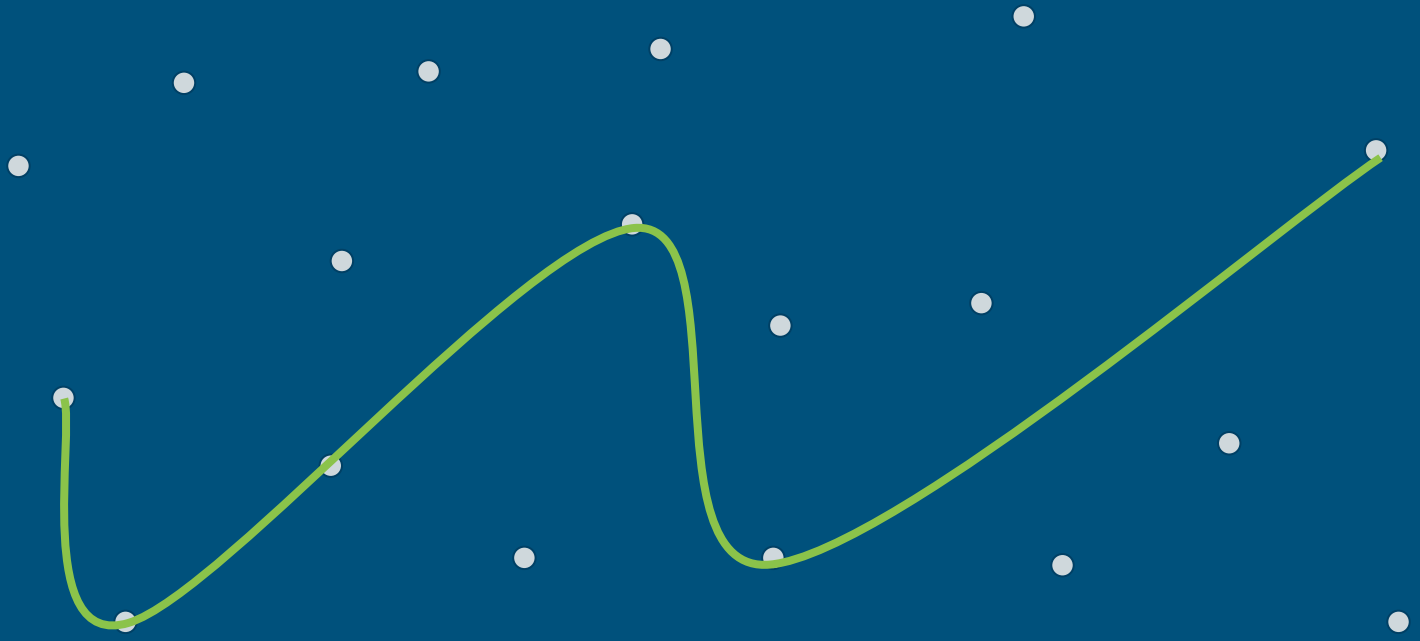
1. Give background and context for systems involved
2. Summarize the incident
3. Work through timeline
    a. As much detail as possible
    b. Reference supporting documents
    c. Ask questions
4. Root causes

Root Causes

Context Error: Alternative Actions



Context Error: Cherry Picking



Context Error: Shopping Bag

Running the meeting

1. Give background and context for systems involved
2. Summarize the incident
3. Work through timeline
    a. As much detail as possible
    b. Reference supporting documents
    c. Ask questions
4. Root causes
5. Lessons Learned
6. Action items

The primary goal of a postmortem is learning - action items are ancillary.

After the Postmortem

Thank You!

@bengerman13

me@bengerman.com

http://slides.bengerman.com