How Did Things Go Right?

RYAN KITCHENS

LEARNING MORE FROM INCIDENTS
SLOs & SLIs

Error Budgets

War Rooms

Incident Reviews

SRE WORLD

@this_hits_home
Failure is so important
Failure is so important that it’s no longer interesting.
Failure is so important that it’s no longer interesting.
Failure is so important that it’s no longer interesting.

Failure is the normal state.
Failure is so important that it’s no longer interesting.

Failure is the normal state.

We’re pretty good @ preventing incidents.
Failure is so important that it’s no longer interesting.

Failure is the normal state.

We’re pretty good @ preventing incidents.

Recovery > prevention.
Failure is so important that it’s no longer interesting.

Failure is the normal state.

We’re pretty good @ preventing incidents.

Recovery > prevention.

The most important thing we can learn is how to build the capacity to encounter failure successfully.
Failure is so important that it’s no longer interesting.

Failure is the normal state.

We’re pretty good @ preventing incidents.

Recovery > prevention.
Failure is so important that it’s no longer interesting. Failure is the normal state. We’re pretty good at preventing incidents. Recovery > prevention. The nines don’t matter. @this.hits_home
Failure is so important that it’s no longer interesting.

Failure is the normal state.

We’re pretty good preventing incidents.

Recovery > prevention.

Availability is made up & the nines don’t matter.
Failure is so important that it’s no longer interesting.

Failure is the normal state.

We’re pretty good @ preventing incidents.

Recovery > prevention.

Availability is made up & the nines don’t matter.

Success isn’t the absence of failure.
Failure is so important that it’s no longer interesting.

Failure is the normal state.

We’re pretty good @ preventing incidents.

Recovery > prevention.

Availability is made up & the nines don’t matter.

Success isn’t the absence of failure.

There is no panacea
Failure is so important that it’s no longer interesting.

We’re pretty good @ preventing incidents.

Recovery > prevention.

Availability is made up & the nines don’t matter.

Success isn’t the absence of failure.

There is no panacea (except for expertise).
There is no root cause.
There is no root cause.
There is no groot cause.
There is no root cause.
There is no root cause.

A PERFECT STORM
There is no A PERFECT STORM root cause.

EACH NECESSARY, BUT ONLY JOINTLY SUFFICIENT
There is no A PERFECT STORM multiple contributing factors. EACH NECESSARY, BUT ONLY JOINTLY SUFFICIENT
SREcon19 Americas Themes

Comprehension  Understandability  Predictability

NONE OF THESE ARE ENTIRELY POSSIBLE

@this_hits_home
NONE OF THESE ARE ENTIRELY POSSIBLE

**SREcon19 Americas Themes**

**Comprehension**

Incidents cannot be fully comprehended.

They are fraught with uncertainty.

Remediation items will contribute to further incidents.

---

NONE OF THESE ARE ENTIRELY POSSIBLE
### SREcon19 Americas Themes

<table>
<thead>
<tr>
<th>Comprehension</th>
<th>Understandability</th>
<th>Predictability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidents cannot be fully comprehended.</td>
<td>Managing the small stuff does not prevent big incidents.</td>
<td></td>
</tr>
<tr>
<td>They are fraught with uncertainty.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remediation items will contribute to further incidents.</td>
<td>Incidents are not made up of causes. We do not find them; we construct them.</td>
<td></td>
</tr>
</tbody>
</table>

NONE OF THESE ARE ENTIRELY POSSIBLE
NONE OF THESE ARE ENTIRELY POSSIBLE

SREcon19 Americas Themes

Comprehension

Incidents cannot be fully comprehended.
They are fraught with uncertainty.
Remediation items will contribute to further incidents.

Understandability

Managing the small stuff does not prevent big incidents.
Incidents are not made up of causes. We do not find them; we construct them.

Predictability

Learning from the last incident will not allow you to predict the next one.
Complex systems are not deterministic.
Their state cannot be precisely or repeatedly foretold.

@this_hits_home
Why do we expect every engineer and operator to be Laplace's Demon?

8:30 AM - 11 Oct 2018
TO LEARN MORE FROM INCIDENTS
Ask more ‘how’ than ‘why’
5 Whys is a competitive disadvantage.
5 Whys is a competitive disadvantage.
There is no root cause.
5 Whys is a competitive disadvantage.
Make sure that things go right

Rather than preventing them from going wrong

SAME PROCESS, DIFFERENT OUTCOMES

SUCCESS

FAILURE
Make sure that things go right

Rather than preventing them from going wrong

SAME PROCESS, DIFFERENT OUTCOMES

@this_hits_home
Make sure that things go right

Rather than preventing them from going wrong

SAME PROCESS, DIFFERENT OUTCOMES
Make sure that things go right

Rather than preventing them from going wrong

SAME PROCESS, DIFFERENT OUTCOMES
Make sure that things go right

Rather than preventing them from going wrong

SAME PROCESS, DIFFERENT OUTCOMES

@this_hits_home
Make sure that things go right

Rather than preventing them from going wrong

SAME PROCESS, DIFFERENT OUTCOMES

@this_hits_home
Make sure that things go right

Rather than preventing them from going wrong

SAME PROCESS, DIFFERENT OUTCOMES
Make sure that things go right

Rather than preventing them from going wrong

SAME PROCESS, DIFFERENT OUTCOMES

@this_hits_home
Constraining ‘PERFORMANCE VARIABILITY’ in order to remove failures will also remove SUCCESSFUL EVERYDAY WORK!
what if the real success was the incidents we made along the way

9:22 PM - 12 Jan 2019
what if the real success was the incidents we made along the way

9:22 PM - 12 Jan 2019
Traditional investigation uses a small portion of the total experience.
Traditional investigation uses a small portion of the total experience.

What about the 99.999% of the time in which things go right?

Figure by Kelvin Genn

@this_hits_home
Traditional investigation uses a small portion of the total experience.

What about the 99.999% of the time in which things go right?

Figure by Kelvin Genn

@this_hits_home
Traditional investigation uses a small portion of the total experience.

What about the 99.999% of the time in which things go right?
Availability is perceived.
Success is often invisible
Success is often invisible

Typical, everyday performance goes unknown & ignored.
HOW DO WE

SUCCESS AND FAILURE
IS NOT REALLY A DUALITY

@this_hits_home
HOW DO WE

Keep normal work normal

SUCCESS AND FAILURE IS NOT REALLY A DUALITY

@this_hits_home
SUCCESS AND FAILURE IS NOT REALLY A DUALITY

HOW DO WE

Keep normal work normal

AND

KEEP

INCIDENTS WEIRD?

SUCCESS AND FAILURE IS NOT REALLY A DUALITY
Signups Impact Triggered By Website Going Cattywampus

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.
PROD ALERT / INC-1806
Signups Impact Triggered By Website Going Cattywampus

DETAILS
Type: Incident
Regions: Global
Severity: Major
Signup Impact?: Yes
Notifications: Yes
Source of Detection: Paging Alert

DESCRIPTION
We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.

ACTIVITY
Status: STABLE
Assignee: Ryan Kitchens
Watchers: 0

DATES
Created: 03/25/2019 1:45 PM
Updated: 03/25/2019 1:50 PM
Incident Start: 03/25/2019 1:40 PM
Incident Detected: 03/25/2019 1:41 PM
Stabilization Time: 03/25/2019 1:48 PM
PROD ALERT / INC-1806
Signups Impact Triggered By Website Going Cattywampus

DETAILS
Type Incident
Regions Global
Severity Major
Signup Impact? Yes
Notifications Yes
Source of Detection Paging Alert

ACTIVITY
Status STABLE
Assignee Ryan Kitchens
Watchers 0

DATES
Created 03/25/2019 1:45 PM
Updated 03/25/2019 1:50 PM
Incident Start 03/25/2019 1:40 PM
Incident Detected 03/25/2019 1:41 PM
Stabilization Time 03/25/2019 1:48 PM

DESCRIPTION
We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.
**PROD ALERT / INC-1806**

**Signups Impact Triggered By Website Going Cattywampus**

---

**DETAILS**

- **Type**: Incident
- **Regions**: Global
- **Severity**: Major
- **Signup Impact?**: Yes
- **Notifications**: Yes
- **Source of Detection**: Paging Alert

---

**ACTIVITY**

- **Status**: STABLE
- **Assignee**: Ryan Kitchens
- **Watchers**: 0

---

**DATES**

- **Created**: 03/25/2019 1:45 PM
- **Updated**: 03/25/2019 1:50 PM
- **Incident Start**: 03/25/2019 1:40 PM
- **Incident Detected**: 03/25/2019 1:41 PM
- **Stabilization Time**: 03/25/2019 1:48 PM

---

**DESCRIPTION**

We were paged for an abnormal RPS trend which kicked off the investigation. **Signup flows were depressed** in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.
PROD ALERT / INC-1806

Signups Impact Triggered By Website Going Cattywampus

**DETAILS**

- **Type**: Incident
- **Regions**: Global
- **Severity**: Major
- **Signup Impact?**: Yes
- **Notifications**: Yes
- **Source of Detection**: Paging Alert

**DESCRIPTION**

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a **death spiral**.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.

**ACTIVITY**

- **Status**: STABLE
- **Assignee**: Ryan Kitchens
- **Watchers**: 0

**DATES**

- **Created**: 03/25/2019 1:45 PM
- **Updated**: 03/25/2019 1:50 PM
- **Incident Start**: 03/25/2019 1:40 PM
- **Incident Detected**: 03/25/2019 1:41 PM
- **Stabilization Time**: 03/25/2019 1:48 PM
PROD ALERT / INC-1806

Signups Impact Triggered By Website Going Cattywampus

**Details**

<table>
<thead>
<tr>
<th>Type</th>
<th>Incident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regions</td>
<td>Global</td>
</tr>
<tr>
<td>Severity</td>
<td>Major</td>
</tr>
<tr>
<td>Signup Impact?</td>
<td>Yes</td>
</tr>
<tr>
<td>Notifications</td>
<td>Yes</td>
</tr>
<tr>
<td>Source of Detection</td>
<td>Paging Alert</td>
</tr>
</tbody>
</table>

**Activity**

<table>
<thead>
<tr>
<th>Status</th>
<th>STABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignee</td>
<td>Ryan Kitchens</td>
</tr>
<tr>
<td>Watchers</td>
<td>0</td>
</tr>
</tbody>
</table>

**Dates**

- Created: 03/25/2019 1:45 PM
- Updated: 03/25/2019 1:50 PM
- Incident Start: 03/25/2019 1:40 PM
- Incident Detected: 03/25/2019 1:41 PM
- Stabilization Time: 03/25/2019 1:48 PM

**Description**

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.
PROD ALERT / INC-1806

Signups Impact Triggered By Website Going Cattywampus

DETAILS

Type: Incident
Regions: Global
Severity: Major
Signup Impact?: Yes
Notifications: Yes
Source of Detection: Paging Alert

ACTIVITY

Status: STABLE
Assignee: Ryan Kitchens
Watchers: 0

DATES

Created: 03/25/2019 1:45 PM
Updated: 03/25/2019 1:50 PM
Incident Start: 03/25/2019 1:40 PM
Incident Detected: 03/25/2019 1:41 PM
Stabilization Time: 03/25/2019 1:48 PM

DESCRIPTION

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.
Contributors & Enablers

Mitigators

Risks

Difficulties in Handling

Follow-up Items

Artifacts

Timeline

References

Open Questions

CONTRIBUTORS & ENABLERS

Mitigators

Risks

Difficulties in Handling

Follow-up Items

Artifacts

Timeline

References

Open Questions

Contributors & Enablers

Mitigators

Risks

Difficulties in Handling

Follow-up Items

Artifacts

Timeline

References

Open Questions
Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions
Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions
**Contributors & Enablers**

Division of labor across teams.
- Usage of deprecated property UI over the recommended one.
- No type safety or constraints were in place for the property.
- The team had confusion about how to revert the property.
- Property parser is overly optimistic of intended context.
- Clearing massive amounts of individual rules is not trivial.
- Afterwards, we still observed short-circuiting on redemption.
- No feedback to users on effects of removing rules en masse.

**Mitigators**

- SR rule to handle broken headers.
- Only regions in peak traffic window observed the issue.
- Website clusters auto-scaled as expected to meet demand.
- An old ASG was able to be repurposed into additional capacity.
- Rate limiting was observed via recently created dashboards.
- New ASG was deployed which patched a bug in the validation.
- All relevant teams were alerted to the issue in under one minute.
- An engineer discovered the property change by searching events.

**Risks**

- An unstaged, global property changed was performed.
- Large volume of redemptions during the holiday window.
- Invalid rules added to redemption workflows.
- Removal of rules contributed to an increase of traffic.
- Validate commands started failing in mid-tier APIs.
- ID validation was not working as expected.
- Recent move to new instance types had unexpected perf.
Use of deprecated UI over the recommended one.

No type safety or property constraints.

Division of labor across teams.

The team had confusion about how to revert.

Property parser is overly optimistic of intended context.

Clearing massive amounts of individual rules is not easy.

Short-circuiting for redemption persisted long after.

No system feedback to users during mass rule removal.
Proxy rules added to handle broken headers.

Only peak traffic regions observed the issue.

Clusters eventually auto-scaled to meet demand.

Old ASG was repurposed into additional capacity.

New ASG deployed to patch a bug in validation.

Rate-limiting observed via recently saved dashboard.

All relevant teams were alerted under 1min to the issue.

A person discovered the change by searching events.
Contributors & Enablers

An unstaged, global property changed was performed.

Large volume of redemptions during holiday window.

Invalid rules added to redemption workflows.

Removal of rules contributed to an increase of traffic.

Validate cmd started failing in mid-tier APIs.

ID validation was not working as expected.

Unexpected perf from upgrade to new instance types.
"I didn’t know it worked like that."

‘Islands of Knowledge’
“I didn’t know it worked like that.”

‘Islands of Knowledge’

“We didn’t make any changes... did we?”
Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions
Maps, Context, and Tribal Knowledge:
On the Structure and Use of Post-Incident Analysis Artifacts in Software Development and Operations

J. Paul Reed | LUND UNIVERSITY
<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:40</td>
<td>CS contacts about SomeProblem begin.</td>
</tr>
<tr>
<td>18:44</td>
<td>First call about the issue is escalated.</td>
</tr>
<tr>
<td>18:54</td>
<td>Confirms there is an increase in CS call volume for the issue.</td>
</tr>
<tr>
<td>19:01</td>
<td>Team tries to reproduce the problem and determines the problem might involve ...</td>
</tr>
<tr>
<td>19:04</td>
<td>Engineer1 from Team escalates the issue.</td>
</tr>
<tr>
<td>19:11</td>
<td>Engineer1 creates TICKET-XXX ...</td>
</tr>
<tr>
<td>19:28</td>
<td>Asks if SomeService is involved in the issue. Engineer1 suggests paging ServiceTeam instead.</td>
</tr>
<tr>
<td>19:33</td>
<td>Engineer2 from ServiceTeam responds.</td>
</tr>
<tr>
<td>19:35</td>
<td>Working theory at this point is that ...</td>
</tr>
</tbody>
</table>
Trade-offs Under Pressure:
HEURISTICS AND OBSERVATIONS OF TEAMS RESOLVING
INTERNET SERVICE OUTAGES

–John Allspaw
What's going on when it seems like nothing is happening?
HOW HARD ARE people working JUST TO KEEP THE SYSTEM HEALTHY?
HOW HARD ARE people working JUST TO KEEP THE SYSTEM HEALTHY?

How do you THINK ABOUT feeding this back UP THE CHAIN?

@this_hits_home
HOW HARD ARE people working JUST TO KEEP THE SYSTEM HEALTHY?

How do you THINK ABOUT feeding this back UP THE CHAIN? Or do you NOT BECAUSE EVERYTHING looks good?

@this_hits_home
THE HARDEST PROBLEM IN TECH

YOU HAVE TO TALK TO PEOPLE!

THERE IS NO GETTING AROUND THIS

@this_hits_home
Interviews

GETTING FROM
‘why did things go wrong?’
TO ‘how did things go right?’
MIND IF I TAKE NOTES?
if you say 'root cause' you are a cop

9:53 AM - 5 Mar 2019
Learning
TEAMS

HOW DO TEAMS ADAPT SUCCESSFULLY?

@this_hits_home
ASK BETTER QUESTIONS

@this_hits_home
Interviews elicit descriptions to construct how we got here. try to convey what the world looked like from their perspective.
How we respond is important.

WE MUST...

Be more careful.

Avoid making mistakes.

Have more discipline.

Show more thoughtfulness.”
How we respond is important.

“WE MUST...”

- Be more careful.
- Avoid making mistakes.
- Have more discipline.
- Show more thoughtfulness.”
How we respond is important.
‘incidents’ are SURPRISES
Takeaways

Recovery > Prevention.

There is no root cause.

Stop reporting on the nines.

Learn how things go right.
Thank you @this_hits_home

http://continuous.wtf