You can’t stop fires with an ambulance

SRECon Asia/Australia ‘18
Piers Chamberlain
Head of Site Reliability Engineering, Xero
Xero - Beautiful Accounting Software

- Founded in 2006 by Rod Drury
- Combined flexibility and convenience of cloud based software was a great success
- Today we are a leading cloud-based accounting platform for small businesses and their advisers
- Over 1.3 million subscribers in more than 180 countries connected to the platform
- >1,200 product updates last year!
A brief history of SRE at Xero

Established June 2016
- As we migrated from a physical datacenter to a cloud provider
- Most of the old operations team transitioned into product groups

Three geographically separated teams
- Today we’re at around 30 SRE engineers
- Like most of you, we’re hiring
Increasing growth and pace of delivery..

1,386,000

Paying subscribers, as at 31 March 2018
increased number of alerts in production
So we simplified alerting

• Started with generic alerts - for example
  ○ # desired actual hosts
  ○ CPU credit burn
  ○ listening for network errors

• We found a good customer concern signal in the rate of hits to our status and error pages
Klaxon

Error Page
GET/error
Event Proxied to SQS
Simple Queue Service
Capture: source ap user agent CORS Origin Custom Cookie
Klaxon Scheduled Lambda
POST event data to Sumo
+ sumologic

User
status.xero.com
GET/status
DataDog updates status.xero.com to maintenance mode
DataDog trigger's Slack Notification

Sumo Alert metric trigger's DataDog workflow
Triggered #40436: Klaxon - Status Page traffic alert DOWN
Assigned: SRE  Service: SRE
Meeting URL: hangouts.google.com  View in: Datadog

Acknowledged #40436: Klaxon - Status Page traffic alert DOWN Jump | by Chris Keogh

Acknowledged #40436: Klaxon - Status Page traffic alert DOWN Jump | by Chris Keogh

Delegated #40435: SRE alert -sumo- IndexOutOfRangeE... Jump | by Chris Keogh

Acknowledged #40435: SRE alert -sumo- IndexOutOfRangeE... Jump | by Sunil Jolly

2:58 PM Chris Keogh @multivac 111 start "increased error rates on Go and Payroll"
2:58 PM multivac Please use the channel #2018-04-17-incident2 for this incident.
To start with
We just created a manual process, as a series of steps and a flowchart.

Multivac
Is now an indispensable component, and allows other to run an incident with some sensible guard-rails.

Then we built some support
For this manual process in the form of a chatbot.

On incident closedown
It produces a postmortem document with links to all the relevant resources.
Detection and Restoration of Service

• Alerting no longer depended on our customer support people noticing a pattern

• Consistent incident process for management enabled us to work the issues without seeding panic

• Good wins in the time to detect and time to resolve (MTTD / MTTR)
But the number of incidents is still a problem
So I went looking for “why”

- I went back over the data and decorated with additional detail
- Themes emerged
  - Similarity in the root cause
  - Often the same teams
  - Often the same deficiencies in detection
Analysis of contributing factors
Our Post-Mortems weren’t working as expected

- We weren’t seeing these trends
- Limited set of participants
- Commitments weren’t communicated
- Work not prioritised
Identified vs. Completed Actions
Finding the influencers

In our organisation these groups were:

1. Tech Group leads
2. Product Owners / Managers
3. Senior Management
Conversation Starters

Service Level Objectives
Even imperfect measurements were opening eyes

Operational Maturity
Especially with new technology

Escalation Frequency
SRE support (or not)?

Release Process
What are compromises we’re making?
Report Card

<table>
<thead>
<tr>
<th>SLO: Latency</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>A check to see whether this service is meeting its server-side response time target.</td>
<td></td>
</tr>
<tr>
<td>WELL DONE! You're operating below the latency threshold of 0.25, as defined in Xero.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SLO: Thresholds Defined</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>A check to see whether SLO's have been defined in SLA Manager.</td>
<td></td>
</tr>
<tr>
<td>WELL DONE! You have defined your error rate and latency.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Xero Appserver</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>We require teams to stay on a whitelisted version of Xero Appserver. Both to cater for AMI vulnerabilities and issues in the Xero Code that runs on top of that AMI.</td>
<td></td>
</tr>
<tr>
<td>THE VERSION OF APPSERVER YOU'RE USING IS 103 DAYS OLD, AND NEW RELEASES OF</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SLO: Error rate</th>
<th>100</th>
</tr>
</thead>
</table>

Update Time: 2018-05-08T05:00:19;
Track Incident Trends

Devise a system to track contributing causes and themes

For us they centred around:

- Teams involved
- Specific Technologies
- Detection Gaps
- Internal / 3rd Party hard dependencies

Key Learning #1

Things I wish we’d done from the start
Key Learning #2
Things I wish we’d done from the start

Involving the right people:
- You need sponsors and leaders.
- Remedial and preventive work won’t happen without their support.
Key Learning #3
Things I wish we’d done from the start

Convince and influence using the right language

- Being the “right” thing to do is not enough
- Use arguments based on
  - customer impact
  - engineering cost
Questions?