How Bad is Your Toil?

Measuring the Human Impact of Process

Kurt Andersen
LinkedIn Product Site Reliability
@drkurta
What is this “toil” of which you speak?
Toil is:

- Manual
- Repetitive
- Automatable
- Tactical
- Without enduring value
- $O(n)$ with service growth (or worse!)
Mean Time To Sleep (MTTS)

- Ryan Frantz & Laurie Denness from Etsy (at the time)
- Velocity 2014
- Looked at the impact of oncall vs. life
- https://github.com/etsy/opsweekly
Round 1 — Time Usage Survey
General Time Categories

- **Toil / Reactive**: Manual, repetitive, lacks enduring value
- **Overhead**: Team meetings, processes, training, admin tasks, hiring
- **S/W Eng**: Writing/maintaining code and associated design & docs
- **System Eng / Architecture**: Making lasting changes to entire systems; consulting on design & architecture
- **Vacation (DTO)**: Taking vacation should not count against other baselines
SRE-Time Survey

- Run monthly from June-November 2018
- Average time to complete: 3m
- Some of the comments were the most instructive
SRE-Time
Demographics

@drkurta
Reactive Work
Reactive work “is the kind of work tied to running a production service that tends to be manual, repetitive, automatable, tactical, devoid of enduring value, and that scales linearly as a service grows.”

Oncall time (either primary or secondary) is a base level of reactivity.

“[Google] SREs report that their top source of reactive work is interrupts (that is, non-urgent service-related messages and emails). The next leading source is on-call (urgent) response, followed by releases and pushes.”
SRE-Time Matrix

5. Last month, please give an estimate of how much time was spent in each of the following categories. The total time spent should total to 100%. Please see definitions below.

<table>
<thead>
<tr>
<th>Category</th>
<th>0 %</th>
<th>10 %</th>
<th>20 %</th>
<th>30 %</th>
<th>40 %</th>
<th>50 %</th>
<th>60 %</th>
<th>70 %</th>
<th>80 %</th>
<th>90 %</th>
<th>100 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software Engineering</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>System Engineering</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Reactive Work</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Overhead</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>DTO</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
SRE-Time Results

- 20% Systems Eng
- 22% SW Eng
- 37% reactive
- 21% Overhead
[Tool X] needs to go, or be supplemented… Its limitations flow into every corner of how we do our job…

I find that interviewing can be a huge off-the-top hit against potentially productive time…

Reactive work for my team is mostly confined to oncall, but our oncall week is almost 100% reactive for 12+ hours a day.

Most of my toil is interrupt driven work, working around broken tooling, horizontal initiatives, and on call tasks. I also go to meetings.
Interlude

Inferring interrupts from agile velocity

- Manual
- Repetitive
- Automatable
- Tactical
- Without enduring value
- $O(n)$ with service growth

@drkurta
Round 2 —
Oncall Pain Working Group

@drkurta
Oncall Pain Survey

- Identify the problem areas that make oncall painful
- Develop tools and strategies to address those areas
- Survey run twice: Spring and Fall 2017

@drkurta
Oncall Pain “Scoring”

• 1 – Not a pain point through
• 5 – Urgently needs improvement
• or N/A if it does not apply to you/your team
On Call Pain

“General Issues”

- Duration of on-call shifts
- Frequency of on-call shifts
- Sleep-interrupting alerts & issues
- Work/Life balance interruption from weekend or Holiday on-call
- Inability to progress on other projects while on-call
- Difficulties contacting or receiving assistance from product POCs (teammates) or developers
- Difficulties contacting or receiving assistance from other teams within LinkedIn
Oncall Pain
“Alerts”

- Spurious alerts requiring no action
- Challenges finding alert-specific severity information
- Challenges finding useful alert-specific resolution documentation
- Alerts due to problems with a downstream service or resource
- Following & cleaning up email-only alerts (low-level)
- Tracking alerts which need to be tuned
- Tuning and updating alerts

@drkurta
Oncall Pain “Daily Activities”

- Mandatory meetings attended by on-call
- Deployments handled by SREs
- Interruptions for deployment/rollback assistance
- LiX / RB approvals
- Interruptions for “urgent” LiX / RB approvals
- Other dev support-related interruptions (detail below)
- Managing incoming team email
- Triaging new tickets
• Are there other aspects of oncall not listed above which you would rate 4-5 (towards the “urgent/highly stressful” end of the scale)?

• For any items you ranked with 4 or above, please provide any details you'd like to add about how these aspects of on-call adversely impact you, and any thoughts you have about how to improve the situation.

• Have there been any changes to how your team handles oncall within the last ~6 months which you feel have had a lasting positive or negative impact in your oncall experience? If so, please elaborate on them here.
## Downstream Issues

<table>
<thead>
<tr>
<th>Issue</th>
<th>Average</th>
<th>Median</th>
<th># of 4+</th>
<th># of N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alerts due to problems with a downstream service or resource</td>
<td>3.8</td>
<td>4</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>Inability to progress on other projects while on-call</td>
<td>3.6</td>
<td>4</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Spurious alerts requiring no action</td>
<td>3.2</td>
<td>3</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>Challenges finding useful alert-specific resolution documentation</td>
<td>3.1</td>
<td>3</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>Sleep-interrupting alerts &amp; issues</td>
<td>3.1</td>
<td>3</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Tuning and updating alerts</td>
<td>2.9</td>
<td>3</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Challenges finding alert-specific severity information</td>
<td>2.8</td>
<td>3</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>Interruptions for deployment/rollback assistance</td>
<td>2.8</td>
<td>3</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Following &amp; cleaning up email-only alerts (low-level)</td>
<td>2.8</td>
<td>3</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Work/Life balance interruption from weekend or Holiday on-call</td>
<td>2.7</td>
<td>3</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Other dev support-related interruptions (detail below)</td>
<td>2.7</td>
<td>3</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Interruptions for &quot;urgent&quot; LiX / RB approvals</td>
<td>2.7</td>
<td>3</td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>
### Alert Management

<table>
<thead>
<tr>
<th>Issue</th>
<th>Average</th>
<th>Median</th>
<th># of 4+</th>
<th># of N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alerts due to problems with a downstream service or resource</td>
<td>3.8</td>
<td>4</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>Inability to progress on other projects while on-call</td>
<td>3.6</td>
<td>4</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Spurious alerts requiring no action</td>
<td>3.2</td>
<td>3</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>Challenges finding useful alert-specific resolution documentation</td>
<td>3.1</td>
<td>3</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>Sleep-interrupting alerts &amp; issues</td>
<td>3.1</td>
<td>3</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Tuning and updating alerts</td>
<td>2.9</td>
<td>3</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Challenges finding alert-specific severity information</td>
<td>2.8</td>
<td>3</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>Interruptions for deployment/rollback assistance</td>
<td>2.8</td>
<td>3</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Following &amp; cleaning up email-only alerts (low-level)</td>
<td>2.8</td>
<td>3</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Work/Life balance interruption from weekend or Holiday on-call</td>
<td>2.7</td>
<td>3</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Other dev support-related interruptions (detail below)</td>
<td>2.7</td>
<td>3</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Interruptions for &quot;urgent&quot; LiX / RB approvals</td>
<td>2.7</td>
<td>3</td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>
## Interruptions

<table>
<thead>
<tr>
<th>Issue</th>
<th>Average</th>
<th>Median</th>
<th># of 4+</th>
<th># of N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alerts due to problems with a downstream service or resource</td>
<td>3.8</td>
<td>4</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>Inability to progress on other projects while on-call</td>
<td>3.6</td>
<td>4</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Spurious alerts requiring no action</td>
<td>3.2</td>
<td>3</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>Challenges finding useful alert-specific resolution documentation</td>
<td>3.1</td>
<td>3</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>Sleep-interrupting alerts &amp; issues</td>
<td>3.1</td>
<td>3</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Tuning and updating alerts</td>
<td>2.9</td>
<td>3</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Challenges finding alert-specific severity information</td>
<td>2.8</td>
<td>3</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>Interruptions for deployment/rollback assistance</td>
<td>2.8</td>
<td>3</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Following &amp; cleaning up email-only alerts (low-level)</td>
<td>2.8</td>
<td>3</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Work/Life balance interruption from weekend or Holiday on-call</td>
<td>2.7</td>
<td>3</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Other dev support-related interruptions (detail below)</td>
<td>2.7</td>
<td>3</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Interruptions for &quot;urgent&quot; LiX / RB approvals</td>
<td>2.7</td>
<td>3</td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>
### Spring'17 Response Patterns by Team

<table>
<thead>
<tr>
<th>Average</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Methods and Tools

<table>
<thead>
<tr>
<th></th>
<th>Downstream</th>
<th>Alerts</th>
<th>Interrupts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linkedout</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alert Correlation</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oncall Retrospective</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Improve alert notes</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Oncall Scheduling</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Dev Oncall</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Oncall Pain
“Fall Follow-up”
Part 1

- Leverage the oncall retro tool to review alerts / identify those needing improvement
- Change in on-call scheduling (4/3, etc)
- Nurse plan with alert correlation to prevent/delay unneeded alert
- Alert note improvements
• Have there been any changes (including those above) to how your team handles on-call within the time since the last survey (April 2017) which you feel have had a lasting positive or negative impact in your on-call experience? If so, please elaborate on them here.
Improvement Results

- Leverage the oncall retro tool to review alerts / identify those needing improvement
- Change in on-call scheduling (4/3, etc)
- Nurse plan with alert correlation to prevent/delay unneeded alert
- Alert note improvements

Legend:
- Much Better
- Slightly Better
- Unchanged
- Slightly Worse
- Much Worse

@drkurta
Round 3 — Tools
INsomnia

- Goal: create a weekly oncall dashboard for the shift handover so that we can have some data to pursue longer term improvement.
- After loading the data into Elasticsearch, the visualization provided by Kibana was better than expected, so most of the time went into improving the dashboard.
Insomnia
Example
Oncall Retro Input

Incident #2955925 (1 of 24)

When: 2017-06-15 17:50 (US/Pacific)
Alert Name: Avg Latency

Your Input

Was this not effective?

- Downstream
- Poorly Tuned
- Self-Resolved
- Already Known

@drkurta
Oncall Retro Results

**Sleep Interrupt Ratio Over Time**

- Ratio of alerts interrupting sleep vs not.

**Values**

- 100
- 75
- 50
- 25
- 0

- 30. Jul
- 6. Aug
- 13. Aug
- 20. Aug
- 27. Aug

- **Interrupted**
- **Not Interrupted**

**Sleep Interrupts per team**

- Total sleep interruptions broken down by segment. Incidents which fire back-to-back are grouped into 1 interruption.

- growth-sre
- messaging-sre
Oncall Retro Results

Effective Incidents

How effective your Incidents are, as measured by those who receive them.

- Effective: 39% (835)
- Ineffective: 52% (1112)
- Skip: 9% (191)

Ineffective Incidents

For your ineffective Incidents, why do people consider them ineffective.

- Downstream: 11% (107)
- Already Known: 28% (274)
- Poorly Tuned: 30% (293)
- Self-Resolved: 31% (302)

@drkurta
Toil is:

- Manual
- Repetitive
- Automatable
- Tactical
- Without enduring value
- $O(n)$ with service growth (or worse!)
Anti-Toil League
Questions?