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

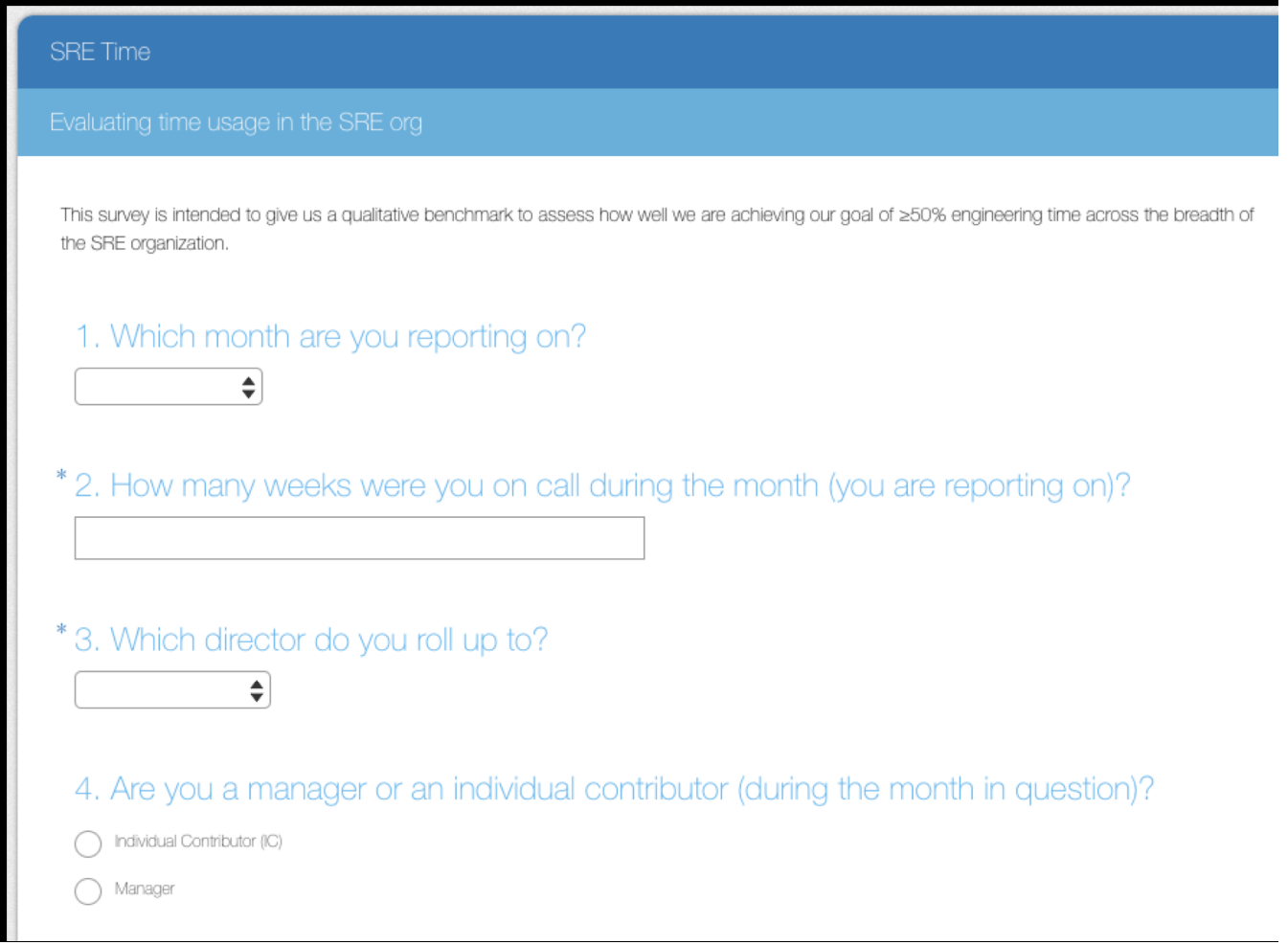
@drkurta

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

Category definitions (from chapter 5 of O'Reilly's *Site Reliability Engineering* book):

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."

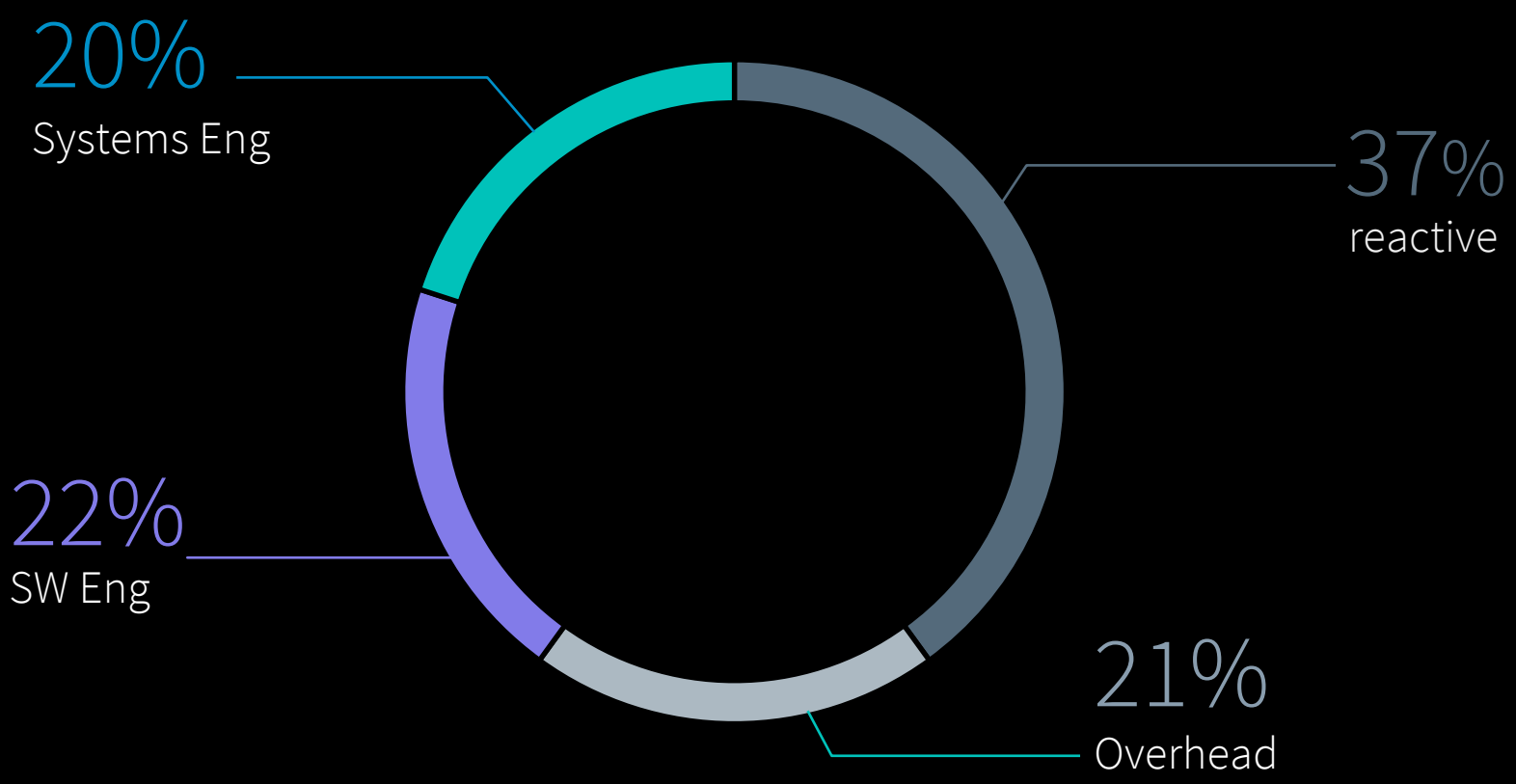
6. Any other comments or questions? Please provide contact info if you want a response.

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.

	0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
Software Engineering	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
System Engineering	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
Reactive Work	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
Overhead	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
DTO	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

SRE-Time Results



@drkurta

SRE-Time Comments

- [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

Round 2 —
Oncall Pain Working Group

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

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

Oncall 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

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

Oncall Pain “Additional Details”

- 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.

Spring'17

Downstream Issues

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

Spring'17

Alert Management

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

@drkurta

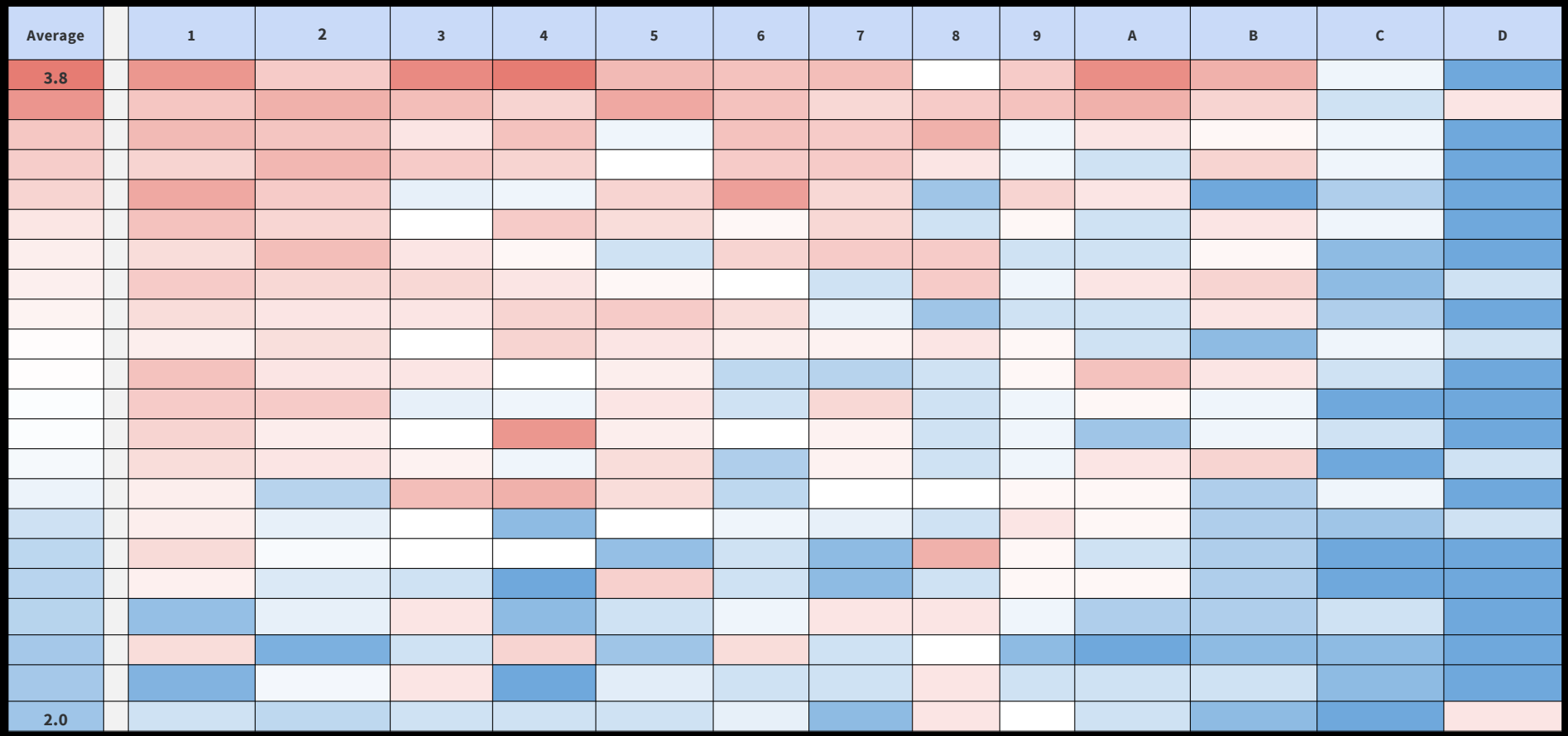
Spring'17

Interruptions

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

Spring'17

Response Patterns by Team



@drkurta

Methods and Tools

	Downstream	Alerts	Interrupts
Linkedout	✓		
Alert Correlation	✓		
Oncall Retrospective	✓	✓	
Improve alert notes		✓	
Oncall Scheduling		✓	✓
Dev Oncall		✓	✓

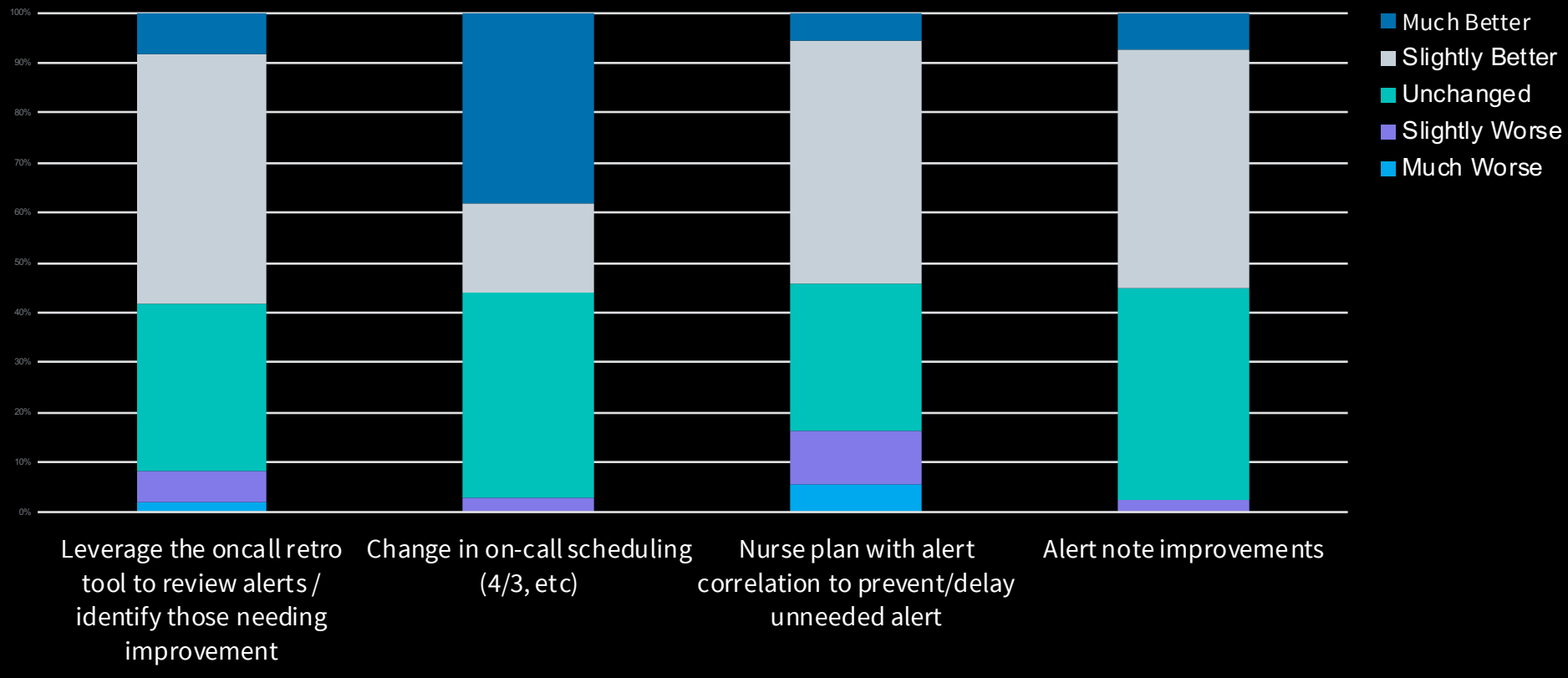
Oncall Pain “Fall Follow-up” Part 1

- 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

Oncall Pain “Fall Follow-up” Part 2

- 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



@drkurta

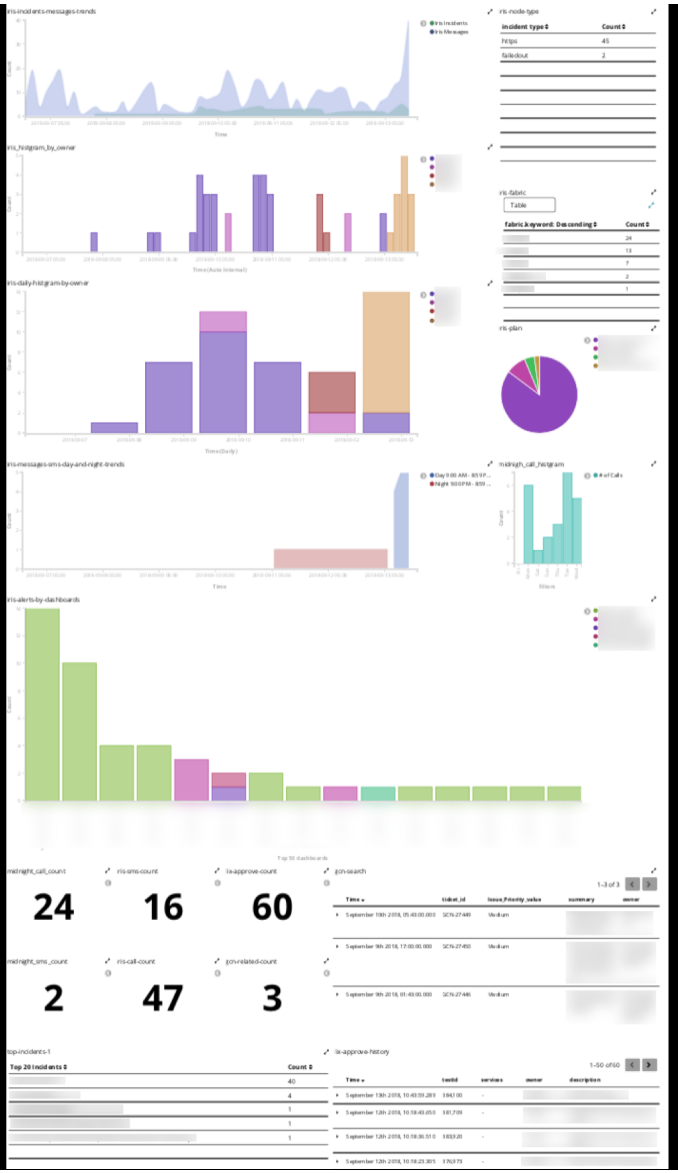
Round 3 — Tools

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

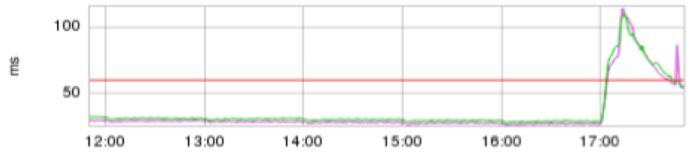


@drkurta

Oncall Retro Input

Incident #2955925 (1 of 24)

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



Your Input

Was was this not effective?

Downsteam | Poorly Tuned | Self-Resolved | Already Known

@drkurta

Oncall Retro Results

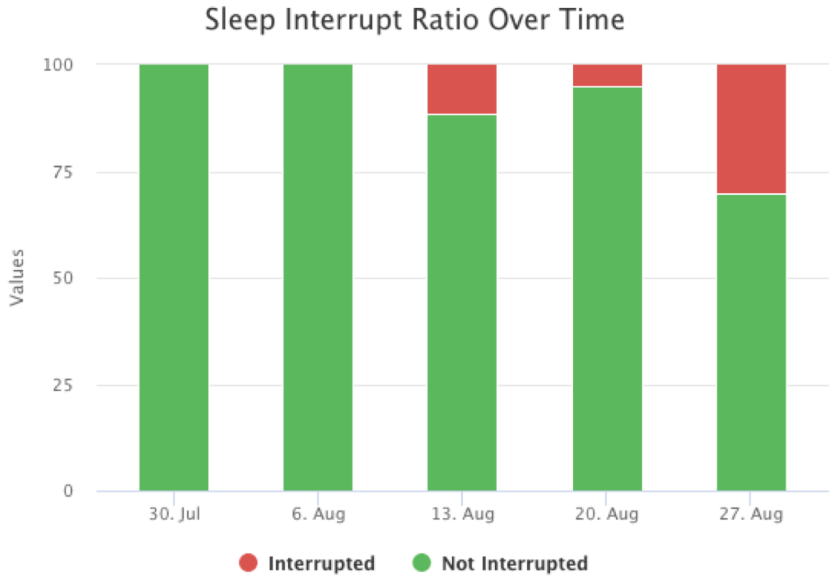
Sleep Interrupts (Aggregate)

Breakdown of S

Not Interrupt

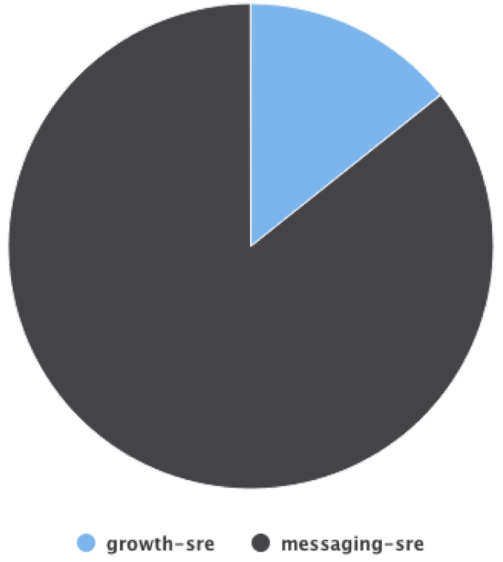
Sleep Interrupt Ratio Over Time

Ratio of alerts interrupting sleep vs not.



Sleep Interrupts per team

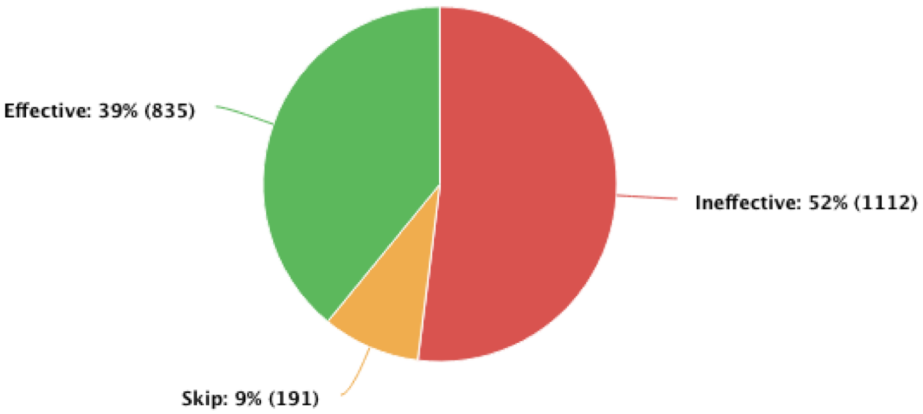
Total sleep interruptions broken down by segment. Incidents which fire back-to-back are grouped into 1 interruption.



Oncall Retro Results

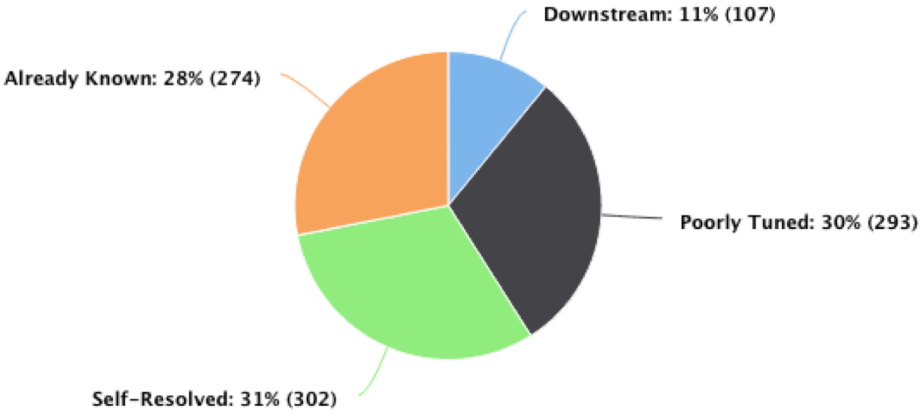
Effective Incidents

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



Ineffective Incidents

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



@drkurta



Toil is:

- Manual
- Repetitive
- Automatable
- Tactical
- Without enduring value
- O(n) with service growth (or worse!)

Anti-Toil League



@drkurta

Questions?

@drkurta


LinkedIn