Matt Brown, @xleem
Customer Reliability Engineer
March, 2018

Know thy enemy

How to prioritize and communicate risk

Matt Brown

Google Cloud

Matt Brown

I'm a kiwi! Live & Work in NZ.

Image: https://pixabay.com/en/new-zealand-island-north-island-309892/, CC0

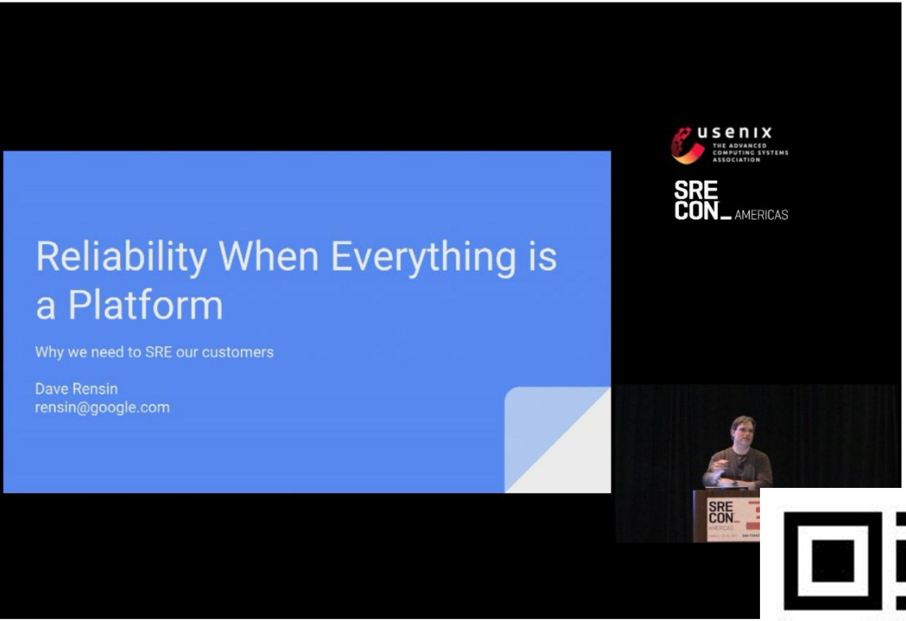
Image: https://pixabay.com/en/new-zealand-island-north-island-309892/, CC0

Matt Brown

I'm a kiwi! Live & Work in NZ.

2nd SREcon, 1st time speaking

Google Cloud



Reliability When Everything is a Platform

Why we need to SRE our customers

Dave Rensin
rensin@google.com

https://goo.gl/T83gcf

Matt Brown

I'm a kiwi! Live & Work in NZ.

2nd SREcon; 1st time speaking

Tech Lead for CRE @ Google

Google Cloud

Agenda

- What is risk?, some observations
- Approaches to risk, why prioritization is needed
- CRE’s first attempt at prioritization
- What Risk Management can teach us about prioritization

What is risk?

a situation involving exposure to danger.

define:risk
google.com

SLO is critical to SRE

SLI

indicator

A measurable quantity representing what's important to users

SLO

objective

The target you want your SLI to reach

SLA

agreement

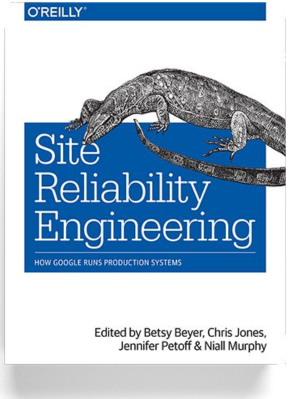
Consequences when the SLO is not met.

Not relevant to today's talk.

Error Budget

1 - SLO

Our primary tool for prioritizing our work.


O'REILLY
Site Reliability Engineering
HOW GOOGLE RUNS PRODUCTION SYSTEMS
Edited by Betsy Beyer, Chris Jones, Jennifer Petoff & Niall Murphy

A situation involving consumption of the error budget

My observations on risk

What’s the biggest risk to your app / service

Image: https://pixabay.com/en/question-mark-why-problem-solution-2123967/, CC0

Image: https://unsplash.com/photos/wS4-XYTyG5k

Many flavours

Personal

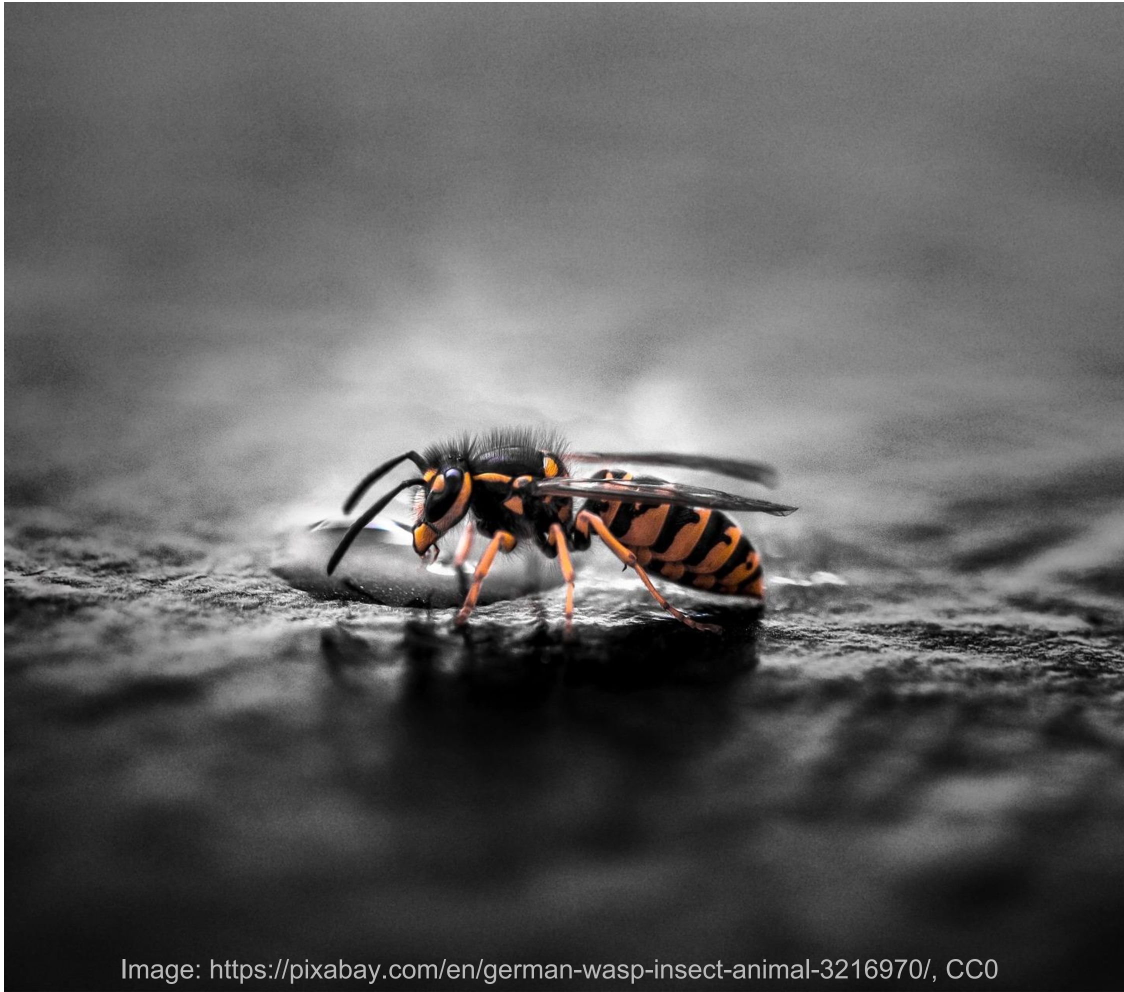


Image: https://pixabay.com/en/german-wasp-insect-animal-3216970/, CC0

Image: https://unsplash.com/photos/wS4-XYTyG5k

Risk can be good

Approaches to risk

Ignorance

Is not bliss

Google Cloud

Image: https://www.pexels.com/photo/beach-wave-948331/, CC0

Image: https://pixabay.com/en/castle-hohenzollern-sunrise-973157/, CC0

Paranoia

Is just as bad

Eliminate

Reduce

Avoid

Image: https://unsplash.com/photos/efc_wvilRs4

Prioritizing risk

Image: https://pixabay.com/en/question-mark-important-sign-1872665/, CC0

Intuition

System
/
Process

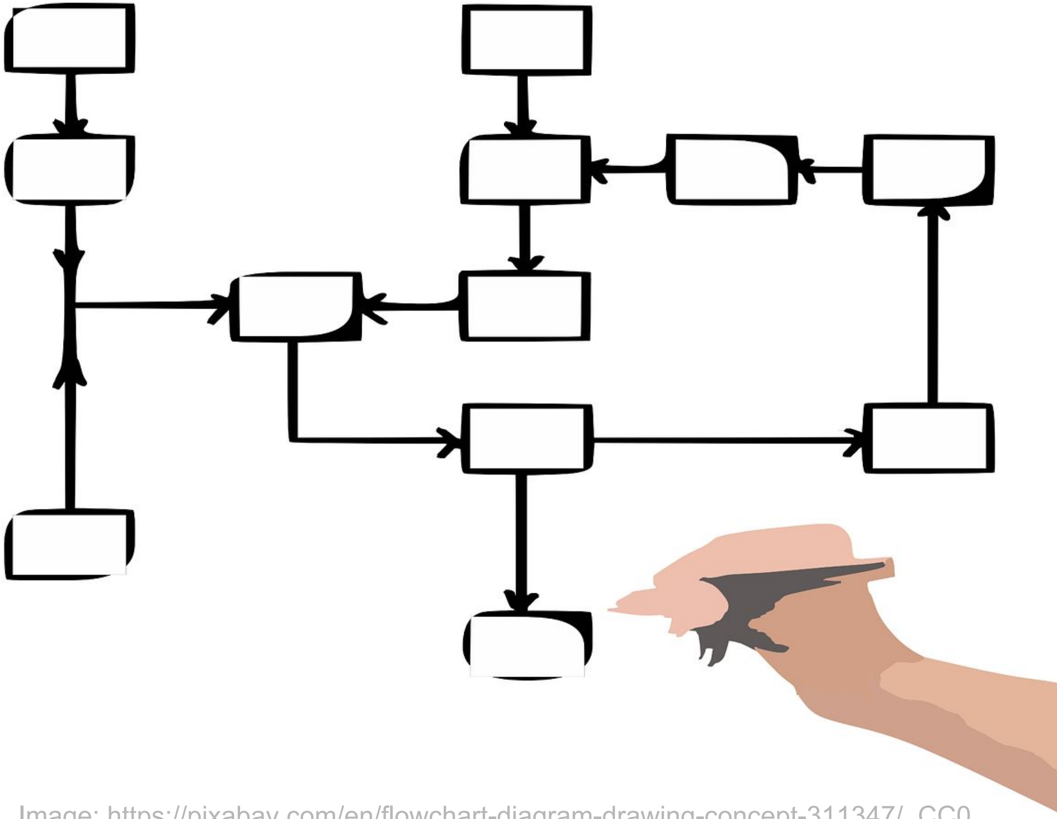


Image: https://pixabay.com/en/flowchart-diagram-drawing-concept-311347/, CC0

Google Cloud

The Risk Matrix

Likelihood

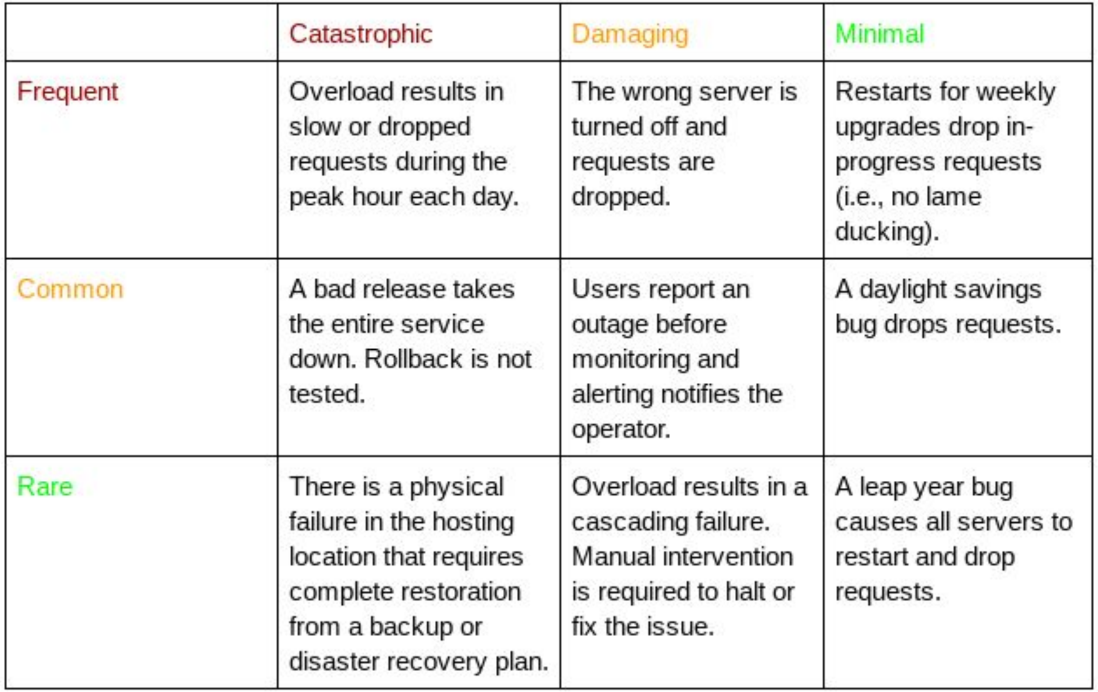
Impact

Images: https://www.pexels.com/photo/white-and-black-dice-37524/ & https://www.pexels.com/photo/time-motion-round-clock-39557/, CC0

The Matrix

Great display, easy to understand

Terrible for prioritization



	Catastrophic	Damaging	Minimal
Frequent	Overload results in slow or dropped requests during the peak hour each day.	The wrong server is turned off and requests are dropped.	Restarts for weekly upgrades drop in-progress requests (i.e., no lame ducking).
Common	A bad release takes the entire service down. Rollback is not tested.	Users report an outage before monitoring and alerting notifies the operator.	A daylight savings bug drops requests.
Rare	There is a physical failure in the hosting location that requires complete restoration from a backup or disaster recovery plan.	Overload results in a cascading failure. Manual intervention is required to halt or fix the issue.	A leap year bug causes all servers to restart and drop requests.

Expected Cost

Expected cost

- Risk Management is a well studied field
- Expected Cost = Probability (Likelihood) * Cost (Impact)
- Costs are easily comparable, solving our matrix problems.
- Can we rephrase our risk characteristics to be able to use this?
- $$ Cost is not always easy for SRE to estimate
- But we already have a budget. A cost is something you spend. We must be able to merge these concepts!

Expected cost for SRE

Likelihood

Quantified as MTBF (days)

Ideally from historical data.

Pragmatically we estimate. (ETBF)

Impact

Quantified as MTTR (typically minutes).

How much of your error budget will this risk consume?

ETTD

ETTR

% Users

Cost

Annual error budget minutes we expect this risk to consume.

Risk Input

Risk Name				
Operator accidentally deletes database; restore from backup required				
Bug in new release breaks uncommon request type				
Physical failure of hosting; implement back-up/DR plan				
Overload causes 15% slow requests at peak each day				
No lame-ducking/health-checks; restarts drop in-flight requests				

https://goo.gl/bnsPj7

Risk Input

Risk Name	ETTD (mins)	ETTR (mins)	% Users	ETBF
Operator accidentally deletes database; restore from backup required	5	480	100	1460
Bug in new release breaks uncommon request type	1440	30	2	90
Physical failure of hosting; implement back-up/DR plan	5	720	100	1095
Overload causes 15% slow requests at peak each day	0	60	15	1
No lame-ducking/health-checks; restarts drop in-flight requests	0	1	100	7

Calculated Expected Cost

Risk Name	ETTD (mins)	ETTR (mins)	% Users	ETBF	**Bad mins/year**
Operator accidentally deletes database	5	480	100	1460	**121**
Bug in new release breaks uncommon request type	1440	30	2	90	**119**
Physical failure of hosting; implement back-up/DR plan	5	720	100	1095	**242**
Overload causes 15% slow requests at peak each day	0	60	15	1	**3287**
No lame-ducking/health-checks; restarts drop requests	0	1	100	7	**52**

Stack Rank

How does this compare to your first guess?

Risk	Bad mins/year
Overload causes 15% slow requests at peak each day	3287
Physical failure of hosting; implement back-up/DR plan	242
Operator accidentally deletes database	121
Bug in new release breaks uncommon request type	119
No lame-ducking/health-checks; restarts drop requests	52

Error budget analysis

Risk	Bad mins/year	99.99%
Overload causes 15% slow requests at peak each day	3287	
Physical failure of hosting; implement back-up/DR plan	242	
Operator accidentally deletes database	121	
Bug in new release breaks uncommon request type	119	
No lame-ducking/health-checks; restarts drop equests	52	

99.99% SLO

52.596 mins/year budget

25% threshold (13.1 mins)

Error budget analysis

Risk	Bad mins/year	99.9%
Overload causes 15% slow requests at peak each day	3287	🟥
Physical failure of hosting; implement back-up/DR plan	242	🟥
Operator accidentally deletes database	121	🟩
Bug in new release breaks uncommon request type	119	🟩
No lame-ducking/health-checks; restarts drop equests	52	🟩

99.9% SLO

525.96 mins/year budget

25% threshold (131 mins)

Risk	Bad mins/year	99.9%
Overload causes 15% slow requests at peak each day	3287	🟥
Physical failure of hosting; implement back-up/DR plan	242	🟥
Operator accidentally deletes database	121	🟨
Bug in new release breaks uncommon request type	119	🟩
...	407	🟩

Error budget analysis

99.9% SLO

525.96 mins/year budget

25% threshold (131 mins)

Takeaways

SLO

You need an SLO, and an error budget.

Foundation for all SRE work and prioritization.

Risks abound

The world is constantly trying to threaten our SLO.

Our job as SREs is to manage that risk.

Prioritization

We can't engage with every risk, we need to prioritize.

Humans are terrible at prioritizing risk.

Estimated Cost

A well established technique for comparing risks.

Breaking a risk into characteristics gives opportunity to reduce bias.

Try it today!

It's easy to apply this technique.

Here's a template spreadsheet you can use: https://goo.gl/bnsPj7

Thank you!

Feedback Welcome

These slides

https://goo.gl/bwT7eC

Me

mattbrown@google.com

@xleem

NEXT 18

g.co/next18

July 24-27, 2018

San Francisco