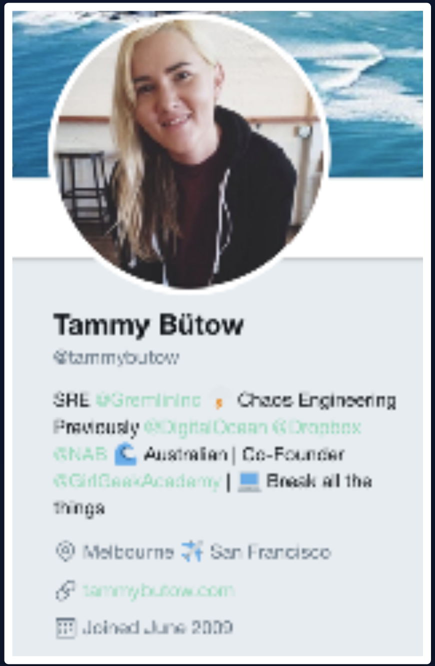
Chaos Engineering:
Why the world needs more resilient systems

@tammybutow

Oh hai, nice to meet you!



Tammy Bütow

@tammybutow

SRE @GremlinInc 🌩 Chaos Engineering Previously @DigitalOcean @Dropbox @NAB 🌊 Australian | Co-Founder @GirlGeekAcademy | 💻 Break all the things

Melbourne ✈️ San Francisco

tammybutow.com

Joined June 2009

Principal SRE @ Gremlin

Tech Advisory Board @ Greenpeace

Enjoys Skateboarding, Snowboarding, Metal, Punk & Breaking Things On Purpose.

@tammybutow

@tammybutow

tammybutow

tb@gremlin.com

Our Gremlin Team Were Previously @

Dropbox

DigitalOcean

National Australia Bank

Queensland University of Technology

PagerDuty

Netflix

Amazon

Salesforce

Google

Datadog

Why the world needs:

More Resilient Systems!

What is a resilient system?

A resilient system is a highly available and durable system.
A resilient system can maintain an acceptable level of service in the face of failure.

A resilient system can weather the storm (a misconfiguration, a large scale natural disaster or controlled chaos engineering).

Let’s review industry examples to understand why we need:

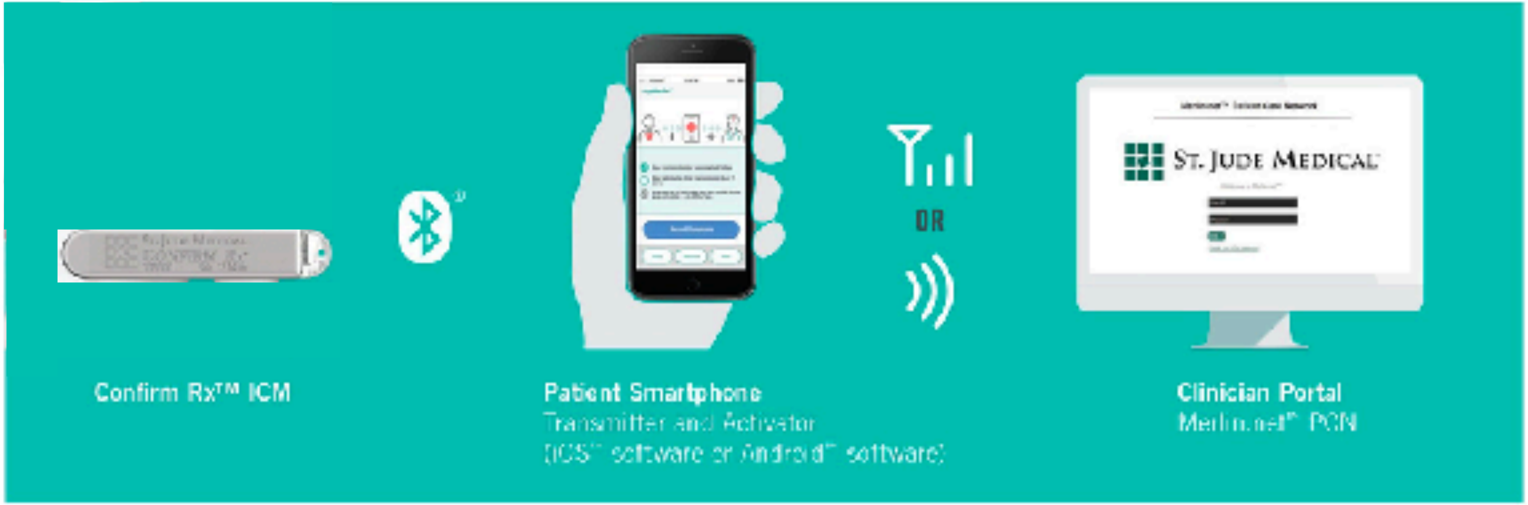
Resilient Systems

Med Tech Industry:

Cardiac monitoring is now done via a
bluetooth device implanted in the body and a mobile app.

The patient takes no action.
Resilience of the device is the only thing the patient cares about.

Confirm Rx™ ICM MONITORING



Confirm Rx™ ICM

Patient Smartphone
Transmitter and Activator
(iOS™ software or Android™ software)

Clinician Portal
Merlin.net™ PCN

myMerlin™ APP FOR PATIENTS ELIMINATES

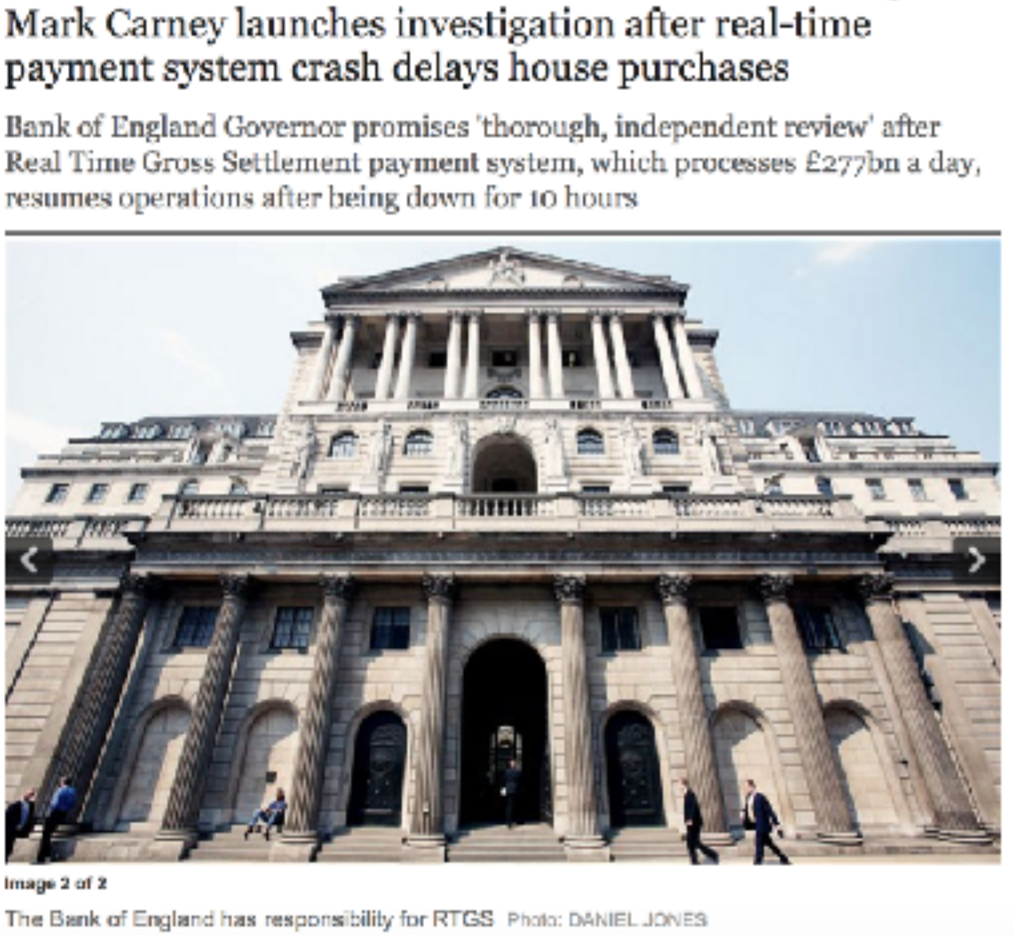
Handheld Activator

Bulky Bedside Transmitter

St. Jude Medical™ mobile transmitters may be purchased for patients without their own mobile device.

Fin Tech Industry:

People are changing jobs, moving homes, traveling and more. Systems need to not only keep up but also provide value anytime/anywhere.



Mark Carney launches investigation after real-time payment system crash delays house purchases

Bank of England Governor promises 'thorough, independent review' after Real Time Gross Settlement payment system, which processes £277bn a day, resumes operations after being down for 10 hours

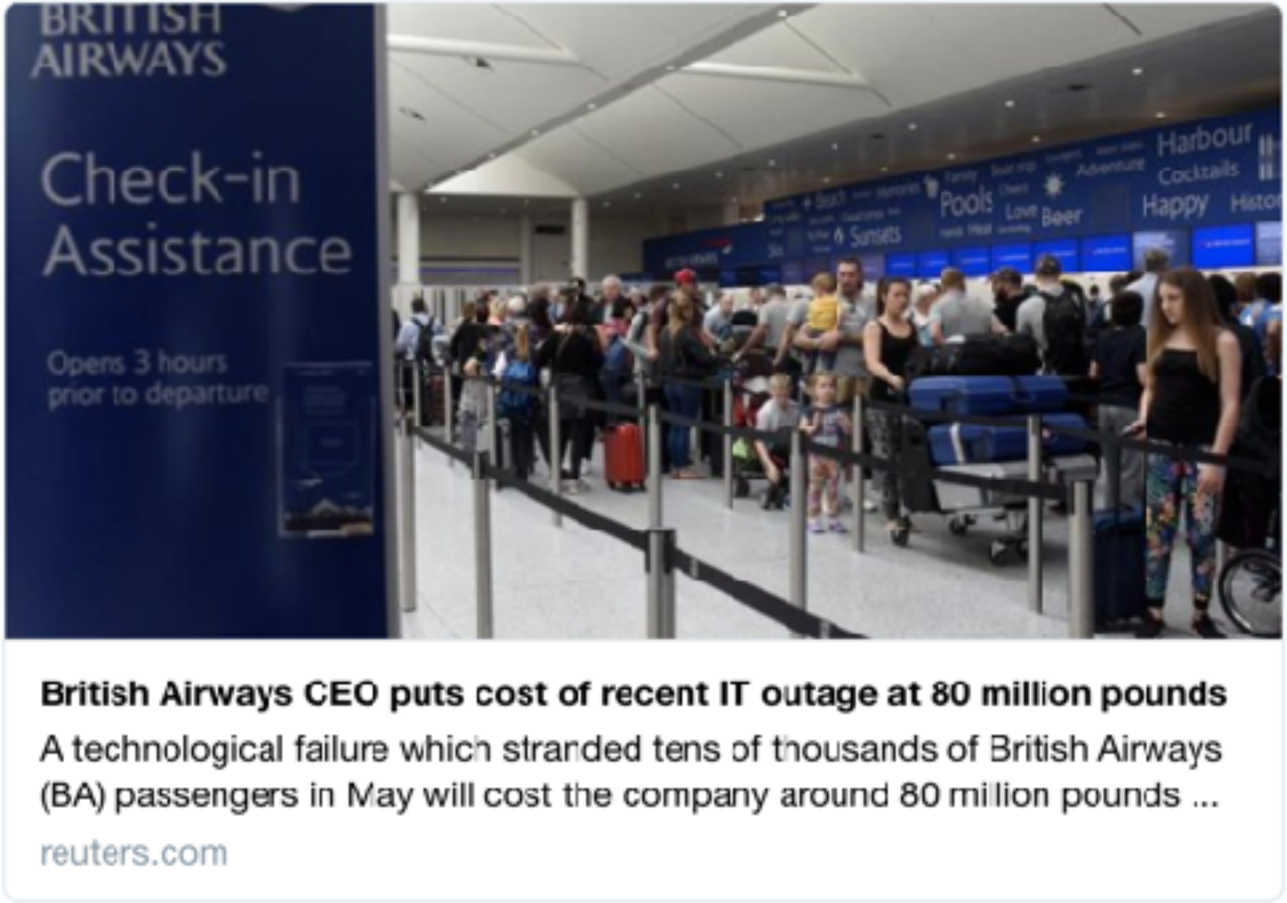
Image 2 of 2

The Bank of England has responsibility for RTGS Photo: DANIEL JONES

A “technical issue related to some routine maintenance”. Impacted the purchase of over 2000 homes.

Transport Tech Industry:

People are traveling so frequently for work and leisure. They need to be able to get where they need to go with no hassles.



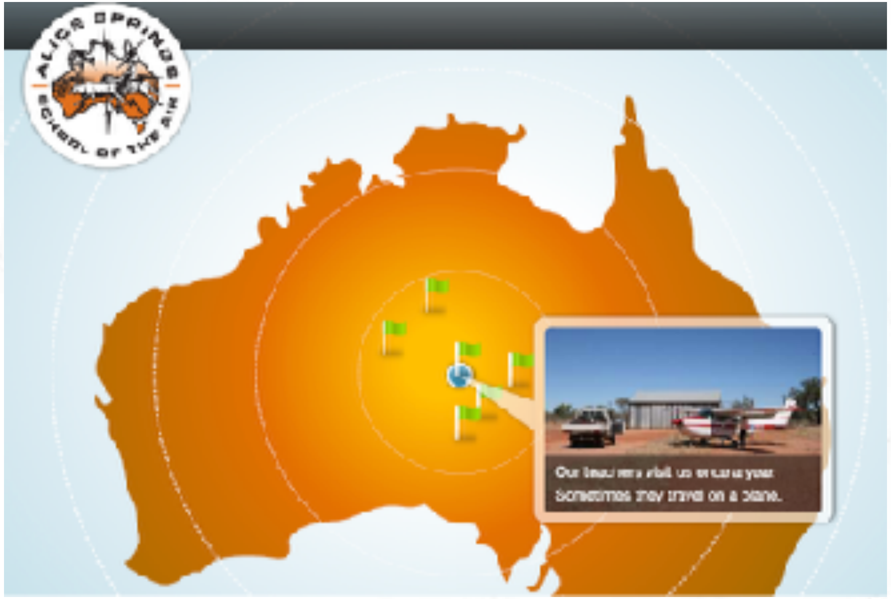
British Airways CEO puts cost of recent IT outage at 80 million pounds

A technological failure which stranded tens of thousands of British Airways (BA) passengers in May will cost the company around 80 million pounds ...

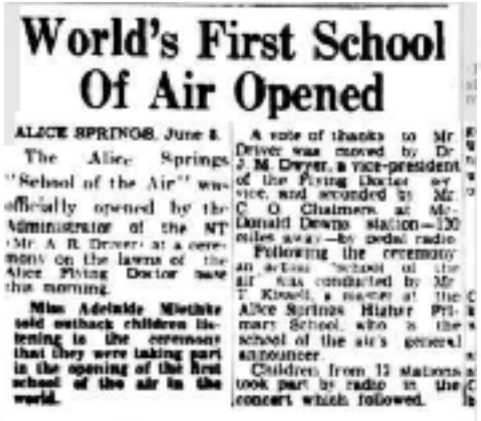
reuters.com

Edu Tech Industry:

More remote learning than ever before. Many students learn remotely. They need reliable access to teachers, students and learning materials.



Our teachers visit us once a year. Sometimes they travel on a plane.



World's First School Of Air Opened

ALICE SPRINGS, June 8.

The Alice Springs "School of the Air" was officially opened by the Administrator of the NT (Mr A. R. Driver) at a ceremony on the lawns of the Alice Flying Doctor base this morning.

Miss Adelaide Miethke told outback children listening to the ceremony that they were taking part in the opening of the first school of the air in the world.

A vote of thanks to Mr Driver was moved by Dr J. M. Dwyer, a vice-president of the Flying Doctor service, and seconded by Mr C. O. Chalmers, at McDonald Downs station—120 miles away—by pedal radio.

Following the ceremony an actual "school of the air" was conducted by Mr T. Kissell, a master at the Alice Springs Higher Primary School, who is the school of the air's general announcer.

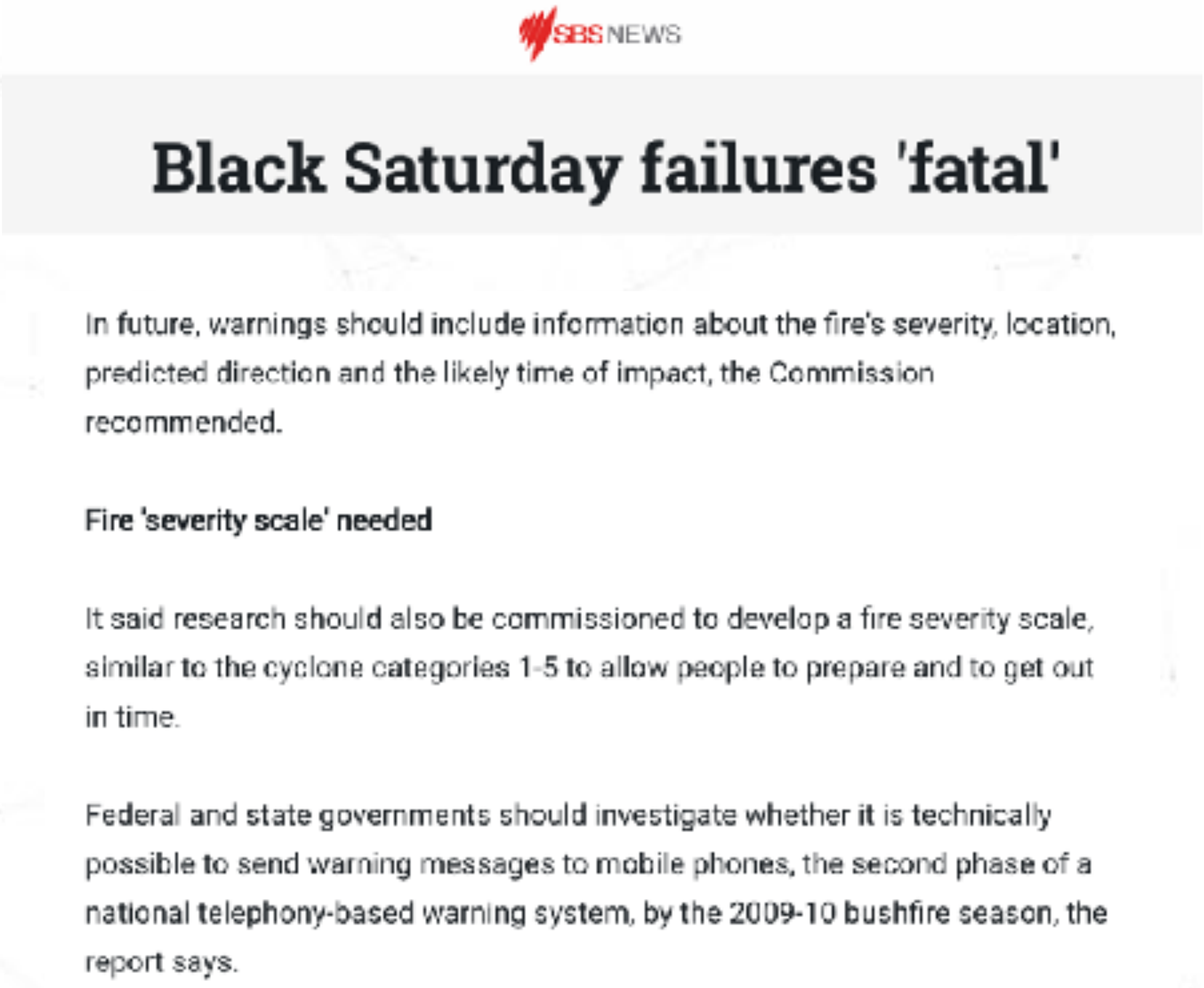
Children from 13 stations took part by radio in the concert which followed.

The world's largest classroom, covering more than 1.3 million square kilometres (502,000 square miles).

Enviro Tech Industry:

People need protection from bushfires, tsunamis, earthquakes and storms. Many of the warning systems for these disasters are legacy unreliable systems.

Saturday, 7 February 2009 - Australia’s all-time worst bushfire disaster



SBS NEWS

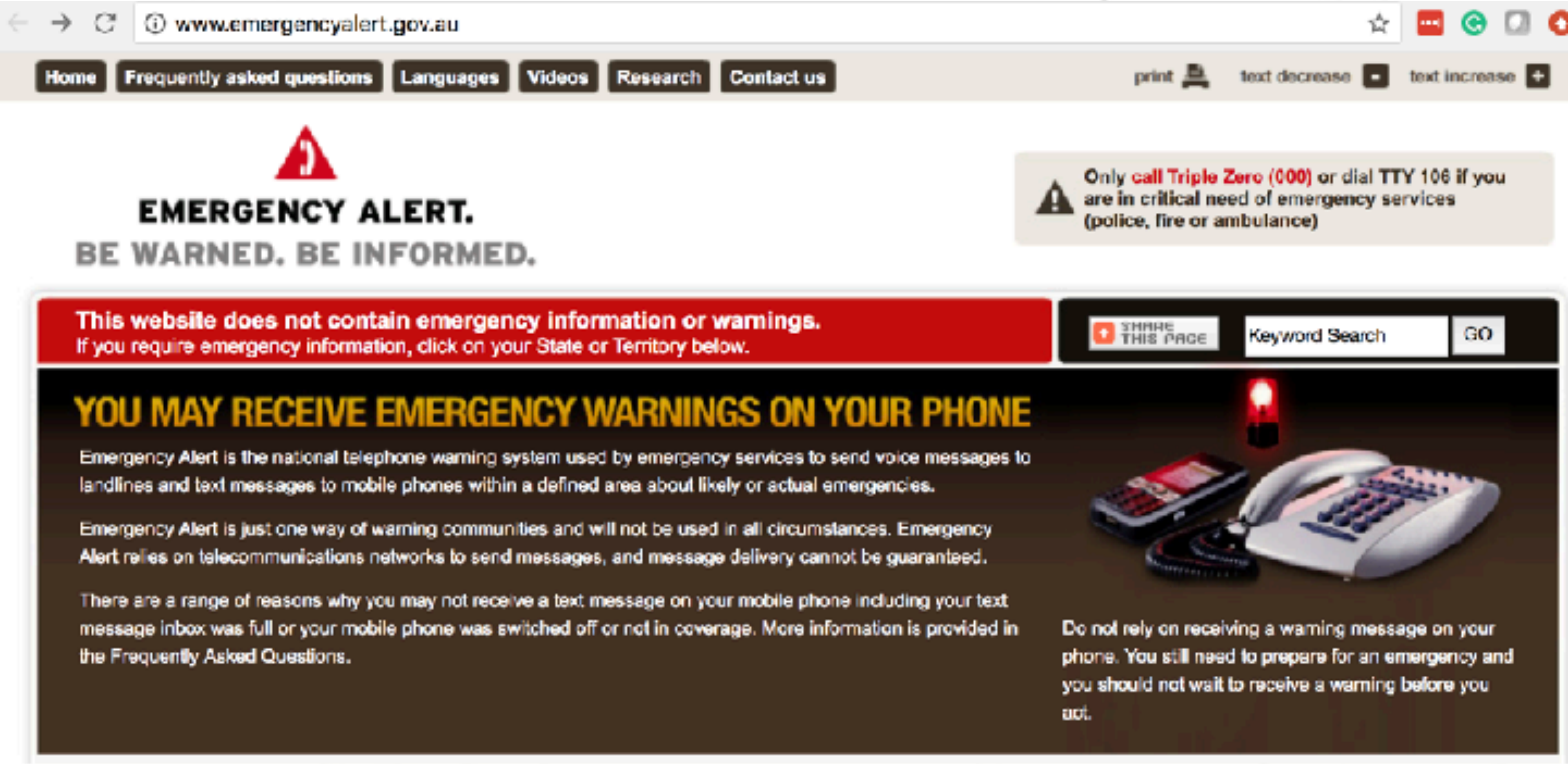
Black Saturday failures 'fatal'

In future, warnings should include information about the fire's severity, location, predicted direction and the likely time of impact, the Commission recommended.

Fire 'severity scale' needed

It said research should also be commissioned to develop a fire severity scale, similar to the cyclone categories 1-5 to allow people to prepare and to get out in time.

Federal and state governments should investigate whether it is technically possible to send warning messages to mobile phones, the second phase of a national telephony-based warning system, by the 2009-10 bushfire season, the report says.



EMERGENCY ALERT.

BE WARNED. BE INFORMED.

Only call Triple Zero (000) or dial TTY 106 if you are in critical need of emergency services (police, fire or ambulance)

This website does not contain emergency information or warnings.
If you require emergency information, click on your State or Territory below.

YOU MAY RECEIVE EMERGENCY WARNINGS ON YOUR PHONE

Emergency Alert is the national telephone warning system used by emergency services to send voice messages to landlines and text messages to mobile phones within a defined area about likely or actual emergencies.

Emergency Alert is just one way of warning communities and will not be used in all circumstances. Emergency Alert relies on telecommunications networks to send messages, and message delivery cannot be guaranteed.

There are a range of reasons why you may not receive a text message on your mobile phone including your text message inbox was full or your mobile phone was switched off or not in coverage. More information is provided in the Frequently Asked Questions.

Do not rely on receiving a warning message on your phone. You still need to prepare for an emergency and you should not wait to receive a warning before you act.

What do these systems have in common?

The primary concern of the user is resilience of the system, in particular high availability.

Let’s figure out how to create:

A great future for everyone

What does a great future look like?

How do we create:

More Resilient Systems?

Introducing:

Chaos Engineering

What is

Chaos Engineering?

Chaos Engineering:

Thoughtful, planned experiments designed to reveal the weakness in our systems.

Inject something harmful, in order to build an immunity

We can inject harm in hosts, containers, pods, applications and more.

What is a

Chaos Engineer?

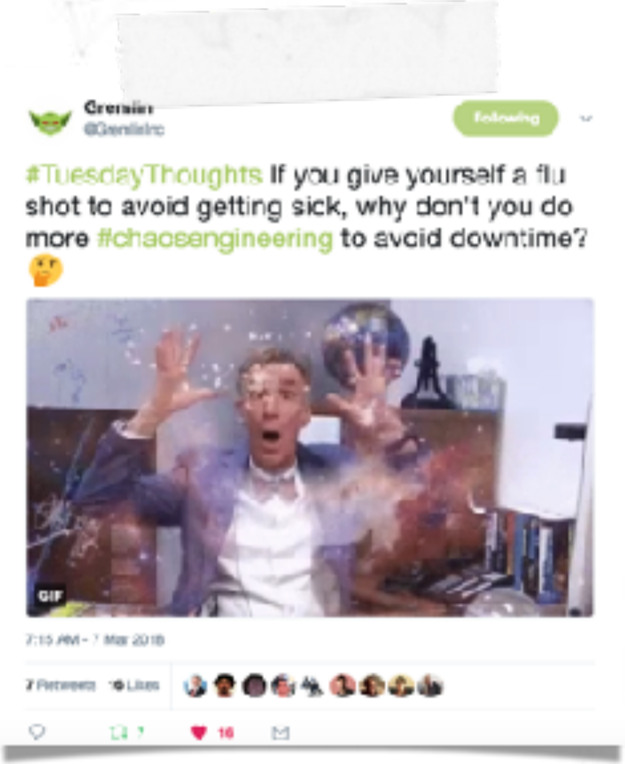
Chaos Engineer:

A vaccine research computer scientist.

SREs / Production Engineers commonly practice Chaos Engineering.

Chaos Engineer:

A vaccine research computer scientist.



Gremlin
@GremlinInc

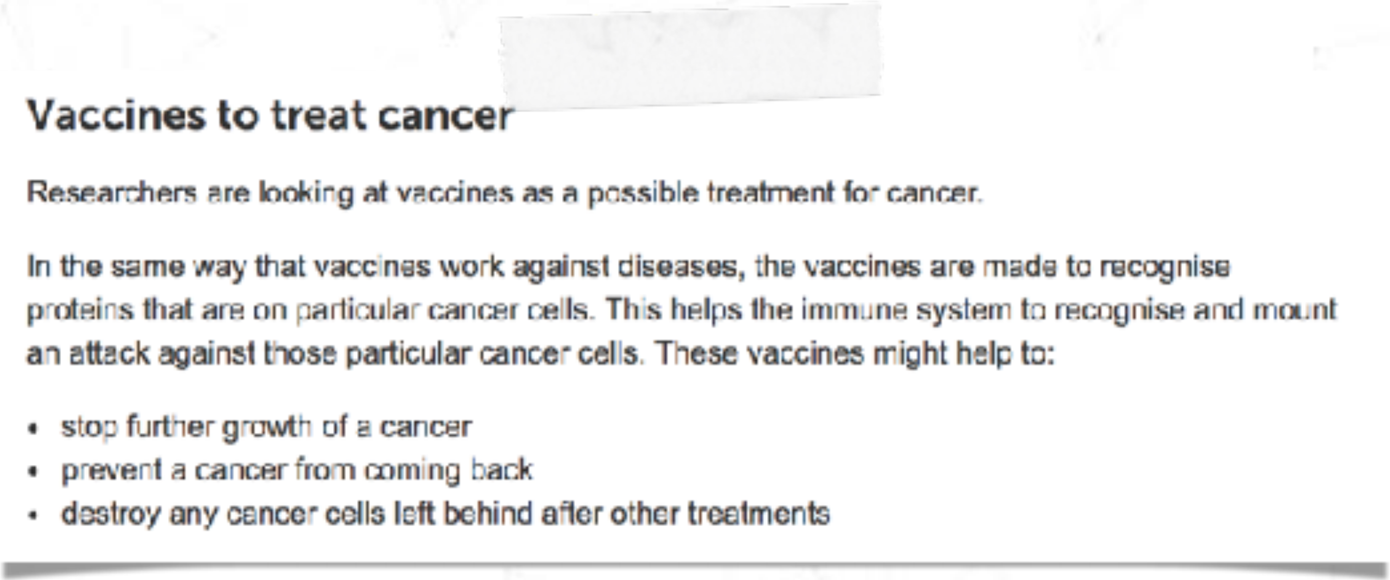
#TuesdayThoughts If you give yourself a flu shot to avoid getting sick, why don't you do more #chaosengineering to avoid downtime? 🤔

7:15 AM - 7 Mar 2018

7 Retweets 16 Likes

Chaos Engineer:

A vaccine research computer scientist.



Vaccines to treat cancer

Researchers are looking at vaccines as a possible treatment for cancer.

In the same way that vaccines work against diseases, the vaccines are made to recognise proteins that are on particular cancer cells. This helps the immune system to recognise and mount an attack against those particular cancer cells. These vaccines might help to:

- stop further growth of a cancer
- prevent a cancer from coming back
- destroy any cancer cells left behind after other treatments

http://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/immunotherapy/types/vaccines-to-treat-cancer

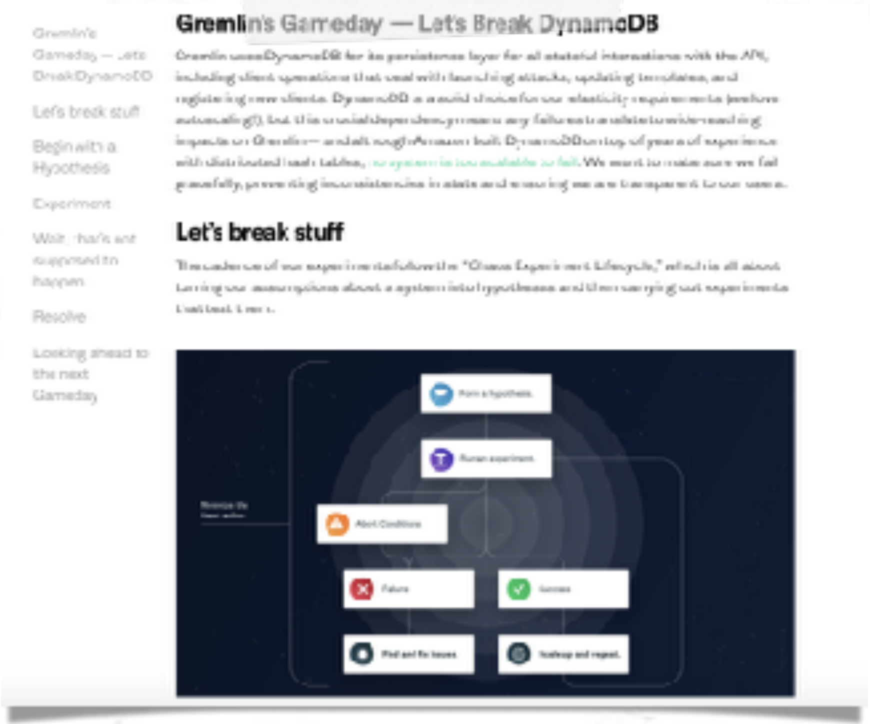
The Bad Database Vaccine

What happens when the database is unreachable?

Does the database fail gracefully?

Does the database have reliable and trustworthy monitoring?

Injecting Harm in DynamoDB



Gremlin's Gameday — Let's Break DynamoDB

Let's break stuff

Begin with a Hypothesis

Experiment

Wait, that's not supposed to happen

Resolve

Looking ahead to the next Gameday

Gremlin's Gameday — Let's Break DynamoDB

Gremlin uses DynamoDB for its persistence layer for all stateful interactions with the API, including client operations that deal with launching attacks, updating templates, and registering new clients. DynamoDB is a solid choice for our elasticity requirements (and autoscaling!), but this crucial dependency means any failures translate to wide-reaching impacts on Gremlin — and although Amazon built DynamoDB on top of years of experience with distributed hash tables, no system is too scalable to fail. We want to make sure we fail gracefully, preventing inconsistencies in state and ensuring we are transparent to our users.

Let's break stuff

The cadence of our experiments follows the "Chaos Experiment Lifecycle," which is all about turning our assumptions about a system into hypotheses and then carrying out experiments that test them.

https://www.gremlin.com/community/tutorials/gremlin-gameday-breaking-dynamodb/

What do you need before you can start doing:

Chaos Engineering

Prerequisites for Chaos Engineering

Prerequisites for Chaos Engineering

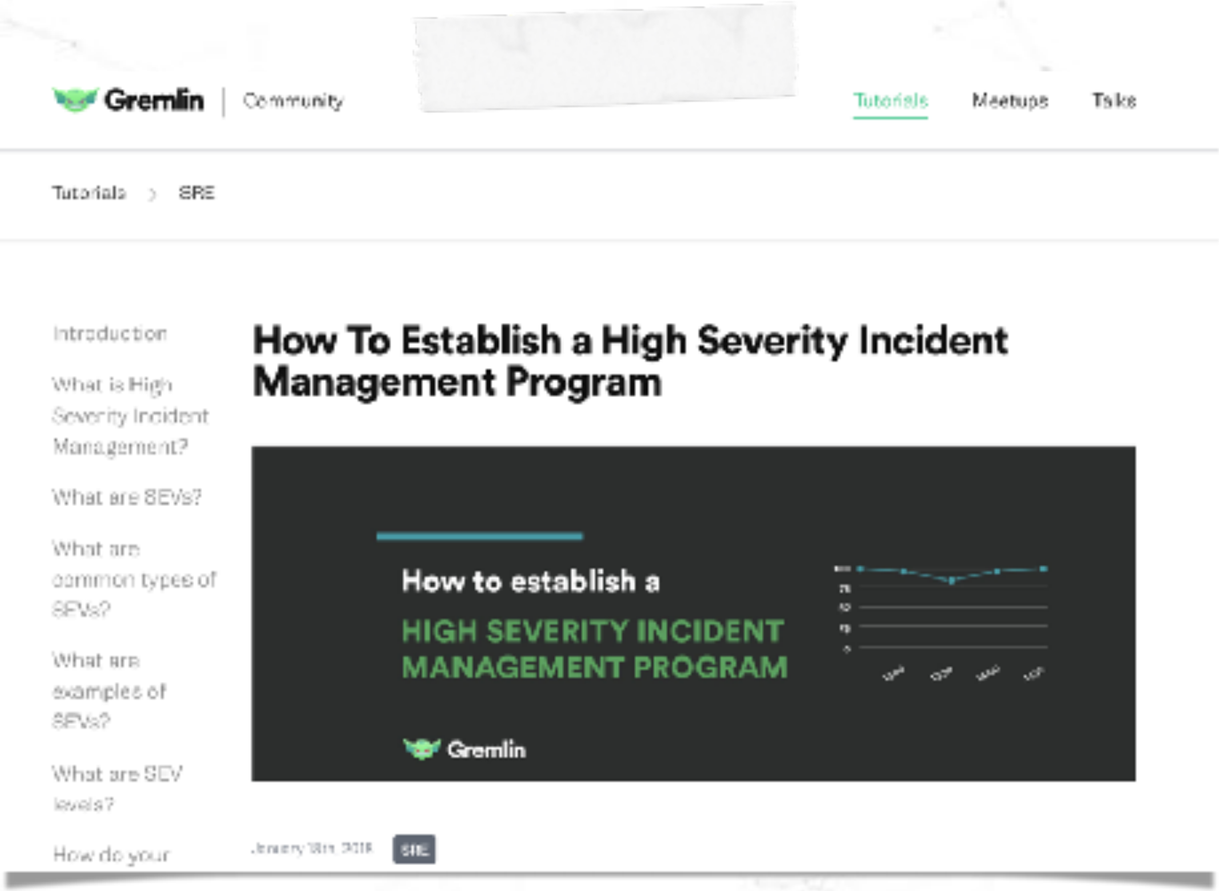
1. High Severity Incident Management
2. Monitoring
3. Measure the Impact of Downtime

Chaos Engineering Prerequisite #1:

High Severity Incident Management

High Severity Incident Management:

The practice of recording, triaging, tracking, and assigning business value to problems that impact critical systems.



Gremlin | Community

Tutorials Meetups Talks

Tutorials > SRE

Introduction

What is High Severity Incident Management?

What are SEVs?

What are common types of SEVs?

What are examples of SEVs?

What are SEV levels?

How do your

How To Establish a High Severity Incident Management Program

January 18th, 2018 SRE

gremlin.com/community

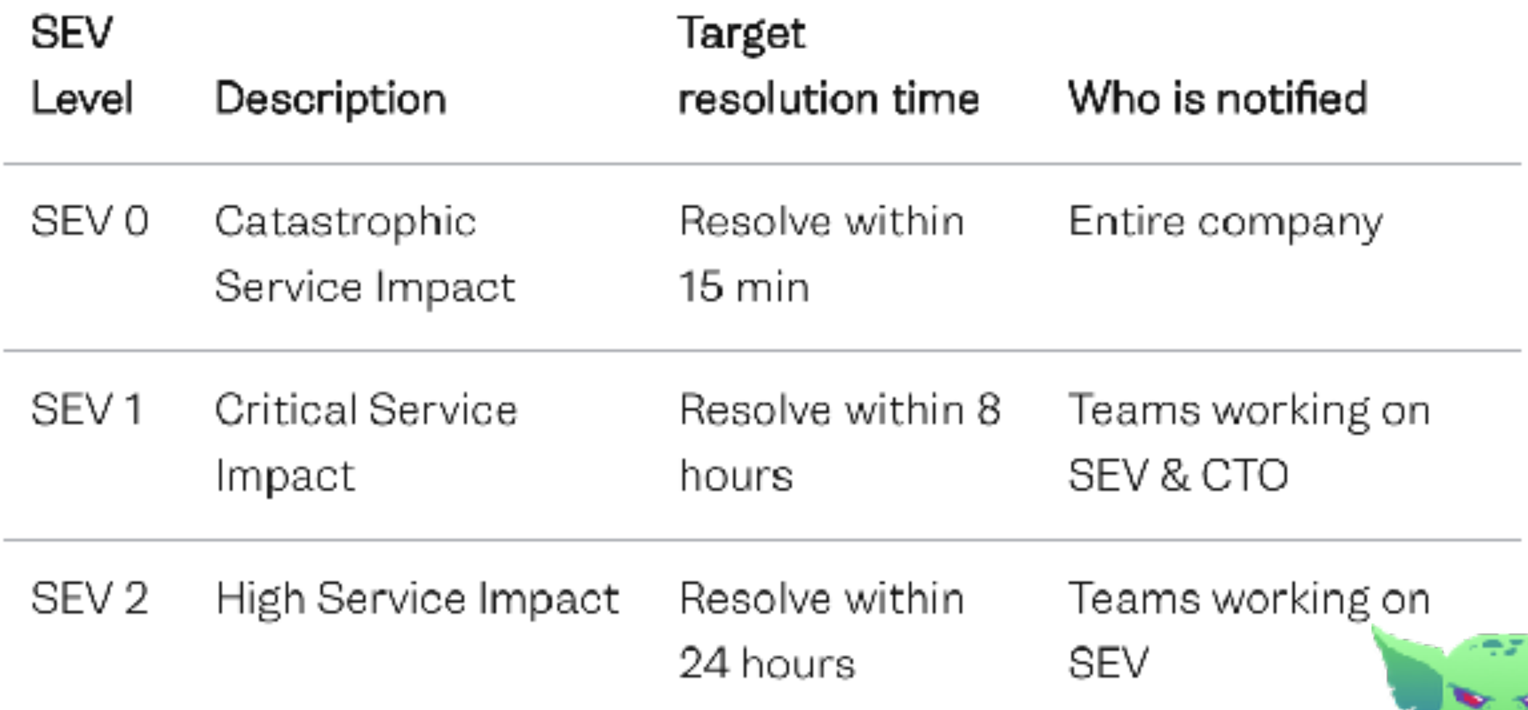
What are

SEVs?

What are SEVs?

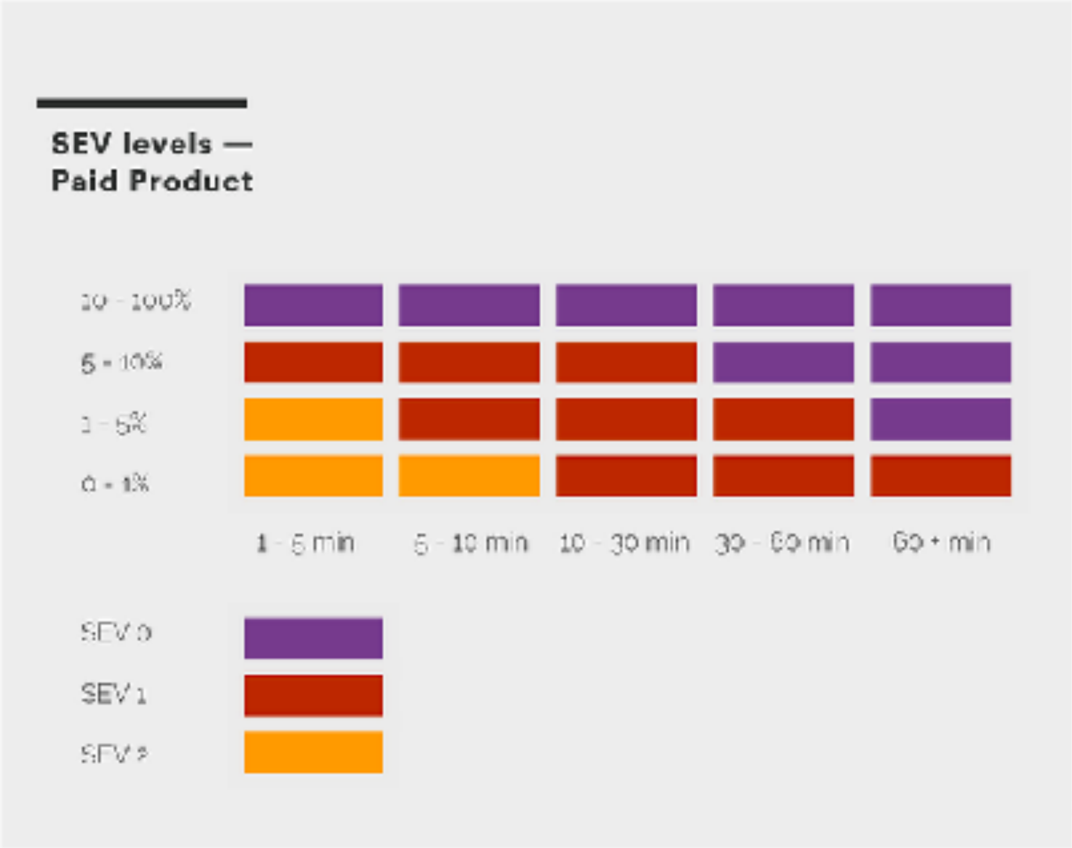
The term SEV is derived from “High Severity Incident”

What are SEVs?



SEV Level	Description	Target resolution time	Who is notified
SEV 0	Catastrophic Service Impact	Resolve within 15 min	Entire company
SEV 1	Critical Service Impact	Resolve within 8 hours	Teams working on SEV & CTO
SEV 2	High Service Impact	Resolve within 24 hours	Teams working on SEV

How Do You Determine SEV levels?



SEV levels — Paid Product

	1 - 5 min	5 - 10 min	10 - 30 min	30 - 60 min	60 + min
10 - 100%	🟪	🟪	🟪	🟪	🟪
5 - 10%	🟥	🟥	🟥	🟪	🟪
1 - 5%	🟧	🟥	🟥	🟥	🟪
0 - 1%	🟧	🟧	🟥	🟥	🟥

SEV 0	🟪
SEV 1	🟥
SEV 2	🟧

What is an example of SEV 0?

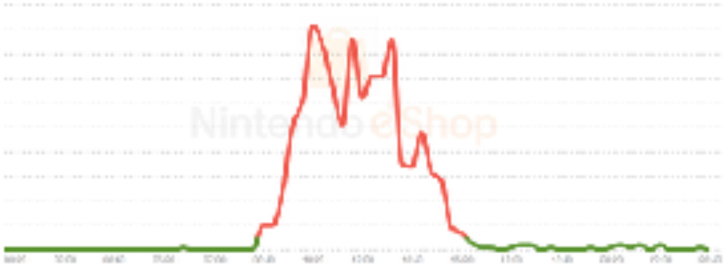
SEV Name: SEV 0 Runaway Cow (auto generated code names help your team remember and refer to SEVs!)

SEV Description: Nintendo Switch eShop is down and not working

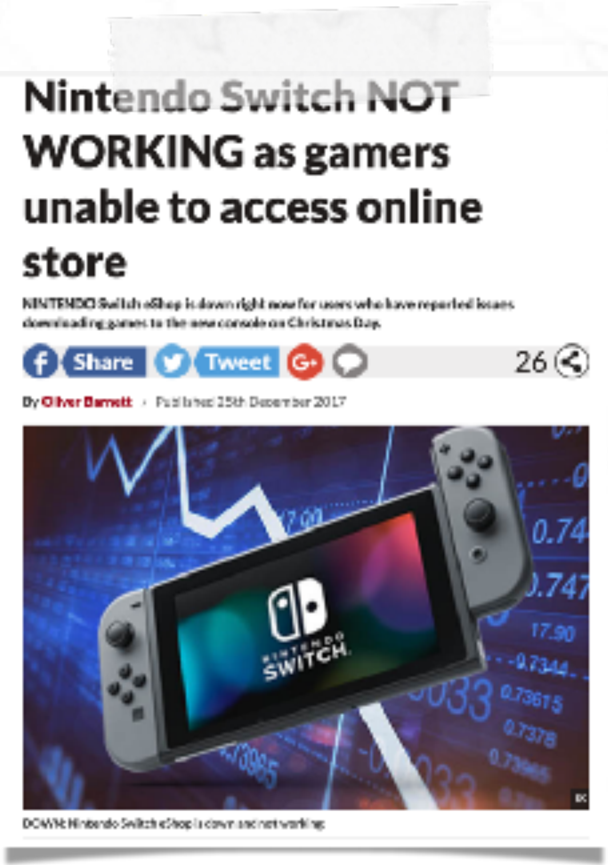
SEV Start Time: 08:40am Dec 25 2017 (Christmas Day)

What is the availability impact? 100%

What is the outage duration? 5 hours and 40 minutes



What is an example of SEV 0?

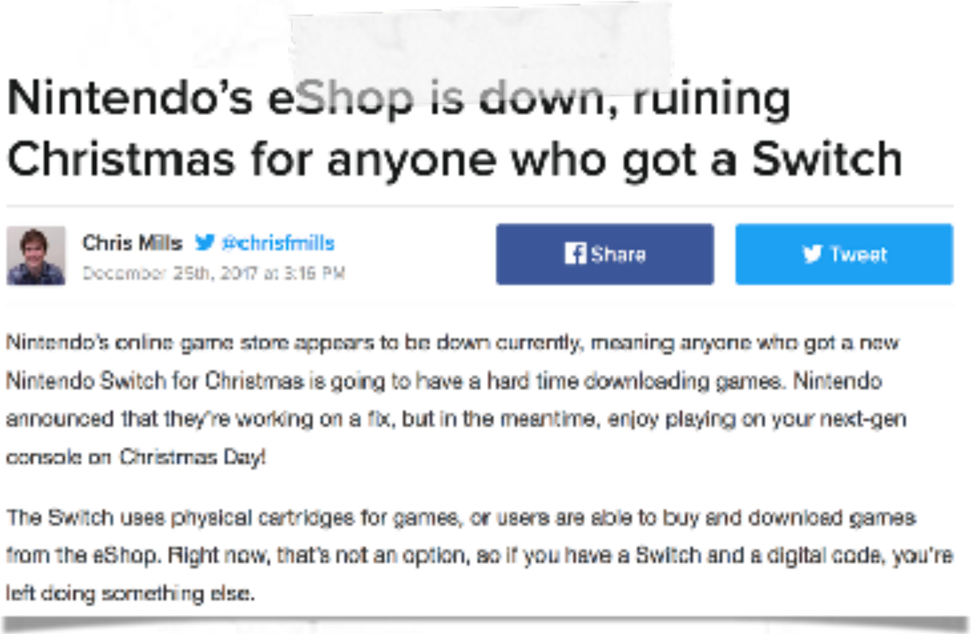


Nintendo Switch NOT WORKING as gamers unable to access online store

NINTENDO Switch eShop is down right now for users who have reported issues downloading games to the new console on Christmas Day.

By Oliver Barnett · Published 25th December 2017

DOWN: Nintendo Switch eShop is down and not working



Nintendo's eShop is down, ruining Christmas for anyone who got a Switch

Chris Mills @chrisfmills
December 25th, 2017 at 3:16 PM

Nintendo's online game store appears to be down currently, meaning anyone who got a new Nintendo Switch for Christmas is going to have a hard time downloading games. Nintendo announced that they're working on a fix, but in the meantime, enjoy playing on your next-gen console on Christmas Day!

The Switch uses physical cartridges for games, or users are able to buy and download games from the eShop. Right now, that's not an option, so if you have a Switch and a digital code, you're left doing something else.

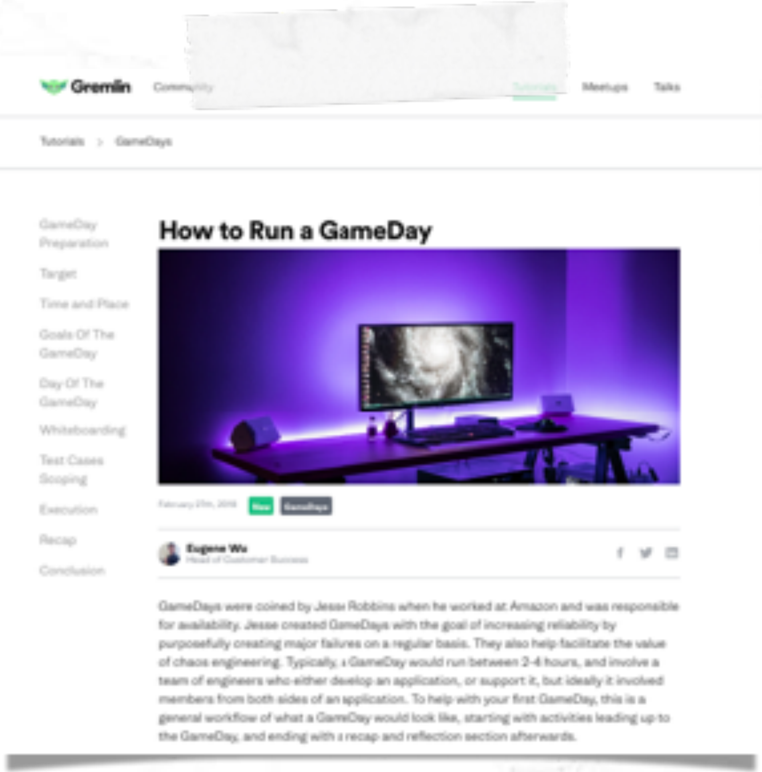
What is the

The SEV Lifecycle?

The SEV Lifecycle

DETECTION	DIAGNOSIS	MITIGATION	PREVENTION	CLOSURE	DETECTION
Alert & page for SEV	Discover source of SEV	Introduce fix and mitigate impact of SEV	Understand root cause and complete all SEV action items	Gameday to replicate SEV and confirm fix is reliable	Alert & page for SEV
TTD (Time to Detection)		TTR (Time to Recovery)	TTP (Time to prevention)		TBF (Time between failures)
TTI (Total time of impact)					

How To Run A GameDay



Gremlin Community Tutorials Meetups Talks

Tutorials > GameDays

GameDay Preparation

Target

Time and Place

Goals Of The GameDay

Day Of The GameDay

Whiteboarding

Test Cases Scoping

Execution

Recap

Conclusion

How to Run a GameDay

February 27th, 2018 New GameDays

Eugene Wu
Head of Customer Success

GameDays were coined by Jesse Robbins when he worked at Amazon and was responsible for availability. Jesse created GameDays with the goal of increasing reliability by purposefully creating major failures on a regular basis. They also help facilitate the value of chaos engineering. Typically, a GameDay would run between 2-4 hours, and involve a team of engineers who either develop an application, or support it, but ideally it involved members from both sides of an application. To help with your first GameDay, this is a general workflow of what a GameDay would look like, starting with activities leading up to the GameDay, and ending with a recap and reflection section afterwards.

gremlin.com/community

How do you identify your critical systems?

What are your critical tier 0 systems?

Traffic
Database
Storage

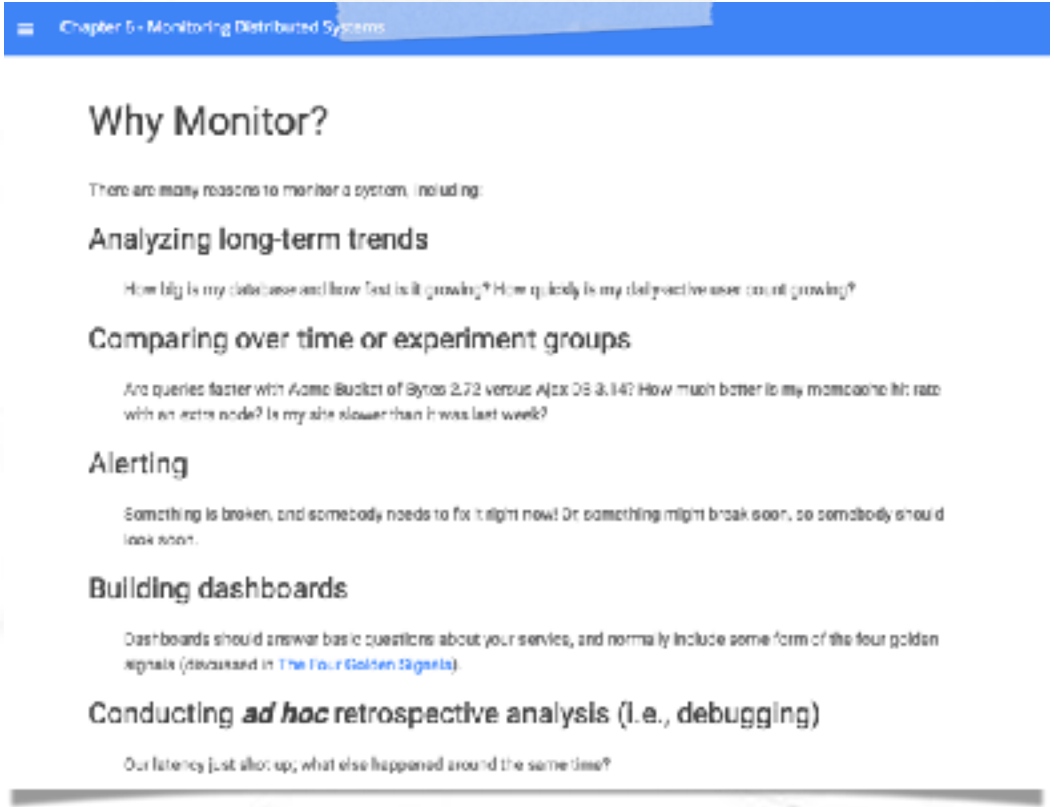
Chaos Engineering Prerequisite #2:

Monitoring

Why Do You Need:

Monitoring

Why Monitor - The Google SRE Book



Chapter 6 - Monitoring Distributed Systems

Why Monitor?

There are many reasons to monitor a system, including:

Analyzing long-term trends

How big is my database and how fast is it growing? How quickly is my daily-active user count growing?

Comparing over time or experiment groups

Are queries faster with Acme Bucket of Bytes 2.72 versus Ajax DB 3.14? How much better is my memcache hit rate with an extra node? Is my site slower than it was last week?

Alerting

Something is broken, and somebody needs to fix it right now! Or, something might break soon, so somebody should look soon.

Building dashboards

Dashboards should answer basic questions about your service, and normally include some form of the four golden signals (discussed in The Four Golden Signals).

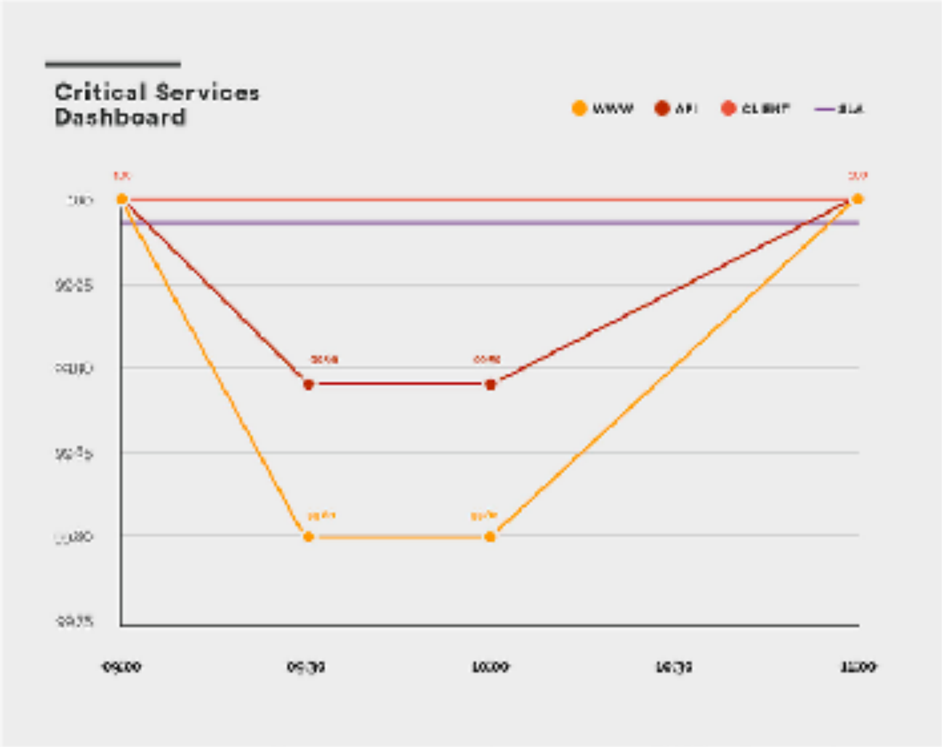
Conducting *ad hoc* retrospective analysis (i.e., debugging)

Our latency just shot up; what else happened around the same time?

https://landing.google.com/sre/book/chapters/monitoring-distributed-systems.html

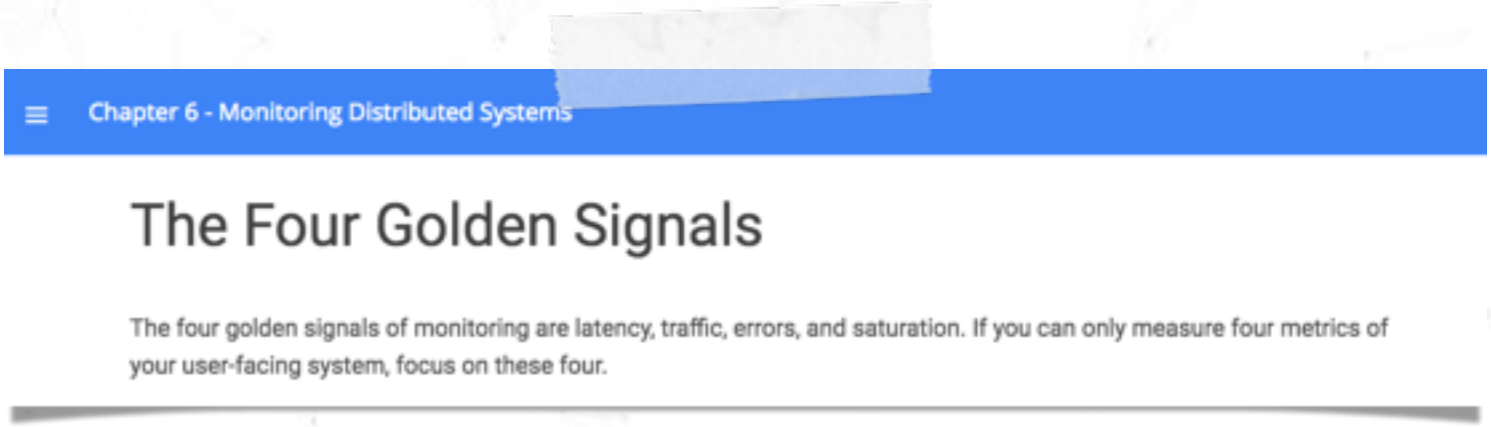
How Should You Use Monitoring

Critical Services Dashboard



gremlin.com/community

The Four Golden Signals - The Google SRE Book



Chapter 6 - Monitoring Distributed Systems

The Four Golden Signals

The four golden signals of monitoring are latency, traffic, errors, and saturation. If you can only measure four metrics of your user-facing system, focus on these four.

https://landing.google.com/sre/book/chapters/monitoring-distributed-systems.html

The Four Golden Signals - The Google SRE Book

Monitoring Signal	Description	Example
Latency	The time it takes to service a request.	HTTP 500 error triggered due to loss of connection to a database
Traffic	A measure of how much demand is being placed on your system	For a web service, this measurement is usually HTTP requests per second
Errors	The rate of requests that fail, either explicitly, implicitly or by policy.	Catching HTTP 500s at your load balancer can do a decent job of catching all completely failed requests.
Saturation	How "full" your service is. Should also signal impending saturation.	It looks like your database will fill its hard drive in 4 hours.

https://landing.google.com/sre/book/chapters/monitoring-distributed-systems.html

What Happens If You Do Chaos Engineering Without Monitoring?

You won’t know what’s happening

Chaos Engineering Prerequisite #3:

Measure The Impact Of Downtime

Measure The Impact Of Downtime

We need to understand how SEV 0s impact our customers and business.

Measure The Impact Of Downtime

System Impact:

- Availability
- Durability

Customer/Business Impact:

- Outcome
- Cost
- Time

What is the impact of the Nintendo Switch eShop SEV 0?

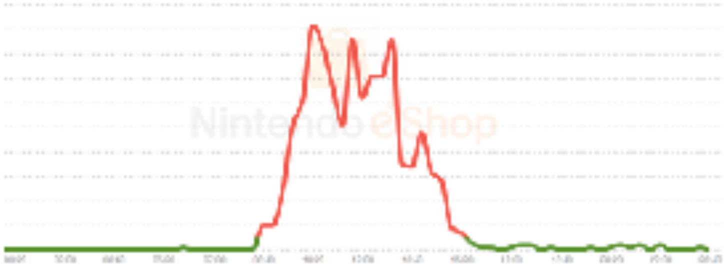
SEV Description: Nintendo Switch eShop is down and not working

What is the availability impact? 100%

Time? 5 hours and 40 minutes

Cost? ______

Outcome? Switch users all over the world can’t buy games



Now we’re ready to get started with:

Chaos Engineering

Chaos Engineering Use Case: Twilio

Chaos Engineering at Twilio with Ratequeue HA

by Michael Wong on November 16, 2017

Twilio engineers are constantly working on improving our core services to meet or exceed 5 nines of availability. A system's capacity for self-healing when a fault occurs is a key measure of achieving high availability. Recently, Twilio used Chaos Engineering to close the gap and eliminate the need for human intervention for common faults involving our core queuing-and-rate-limiting system, Ratequeue.

Power modern communications. Build the next generation of voice and SMS applications.

Posts By Stack

.NET · Arduino · Java · JavaScript · PHP · Python · Ruby · Swift

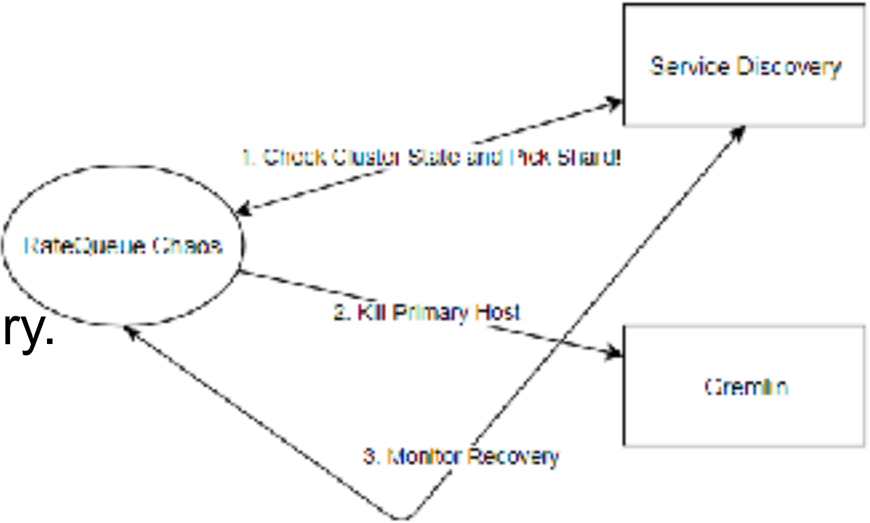
Posts By Product

MMS · Programmable Chat · SIP · SMS · Task Router · Twilio Client · Twilio Video · Voice

Chaos Engineering Case Study: Twilio

Ratequeue Chaos has 3 goals:

1. Pick a shard
2. Kill primary
3. Monitor recovery.



Share The
Chaos Engineering
Journey Widely

Share The Chaos Engineering Journey Widely

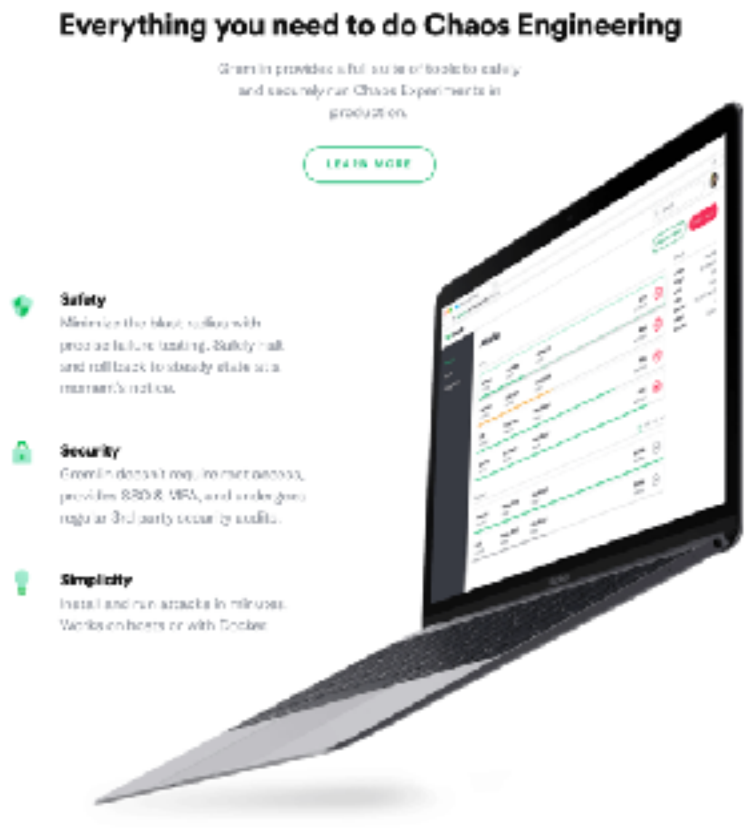
- Do a Chaos Engineering Kick Off @ All Hands
- Send email updates & progress reports
- Run Monthly Metrics Reviews
- Deliver Presentations

Don’t Surprise Everyone!

What is

Gremlin?

What is Gremlin?



Everything you need to do Chaos Engineering

Gremlin provides a full suite of tools to safely and securely run Chaos Experiments in production.

Safety
Minimize the blast radius with precise failure testing. Safely halt and roll back to steady state at a moment's notice.

Security
Gremlin doesn't require internet access, provides SSO & MFA, and undergoes regular 3rd party security audits.

Simplicity
Install and run attacks in minutes. Works on hosts or with Docker.

Gremlin Chaos Engineering Attacks

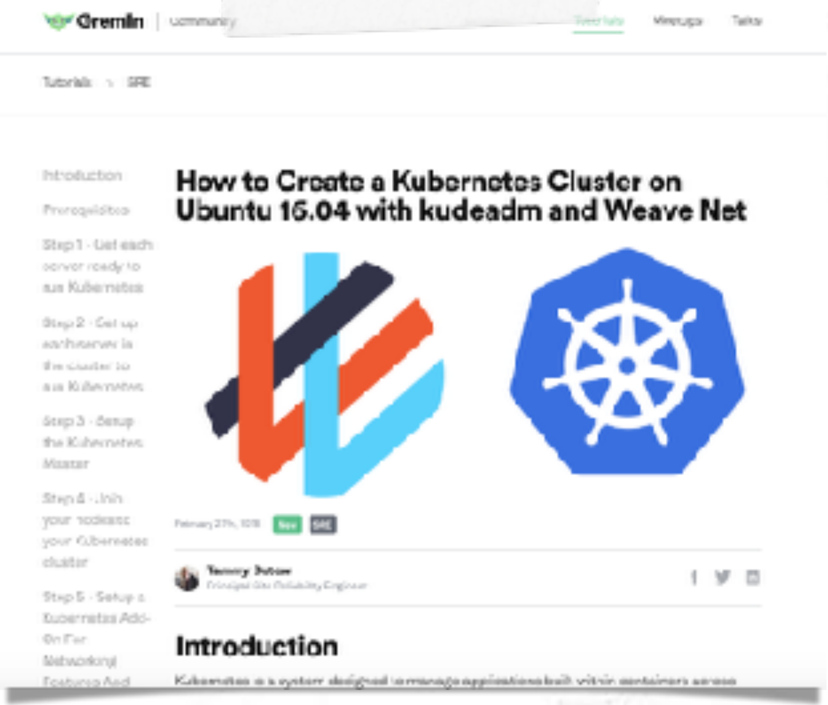
There are a range of attacks built-in and ready to run on Linux.

Type of Attack	**Attack**	**Gremlin Support (March 2018)**
Resource	CPU	✅
Resource	Disk	✅
Resource	IO	✅
Resource	Memory	✅
State	Process Killer	✅
State	Shutdown	✅
State	Time Travel	✅
Network	Blackhole	✅
Network	DNS	✅
Network	Latency	✅
Network	Packet Loss	✅

Live Chaos Engineering

Demo

Create a Kubernetes Cluster



Gremlin | Community

Tutorials Meetups Talks

Tutorials > SRE

Introduction

Prerequisites

Step 1 - Get each server ready to run Kubernetes

Step 2 - Set up each server in the cluster to run Kubernetes

Step 3 - Setup the Kubernetes Master

Step 4 - Join your nodes to your Kubernetes cluster

Step 5 - Setup a Kubernetes Add-On For Networking Features And

How to Create a Kubernetes Cluster on Ubuntu 16.04 with kudeadm and Weave Net

February 27th, 2018 New SRE

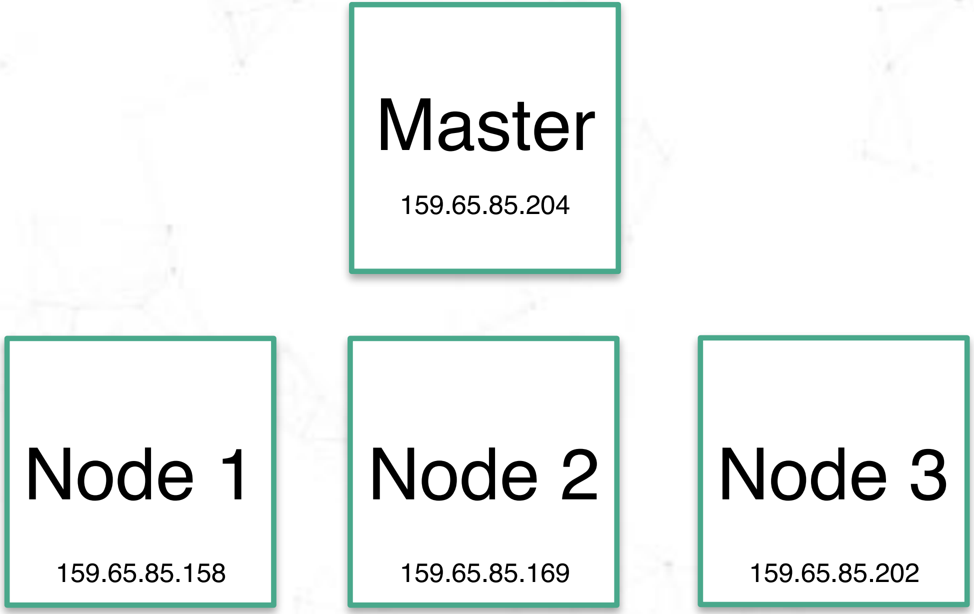
Tammy Butow
Principal Site Reliability Engineer

Introduction

Kubernetes is a system designed to manage applications built within containers across

gremlin.com/community

Create a Kubernetes Cluster



Master

159.65.85.204

Node 1

159.65.85.158

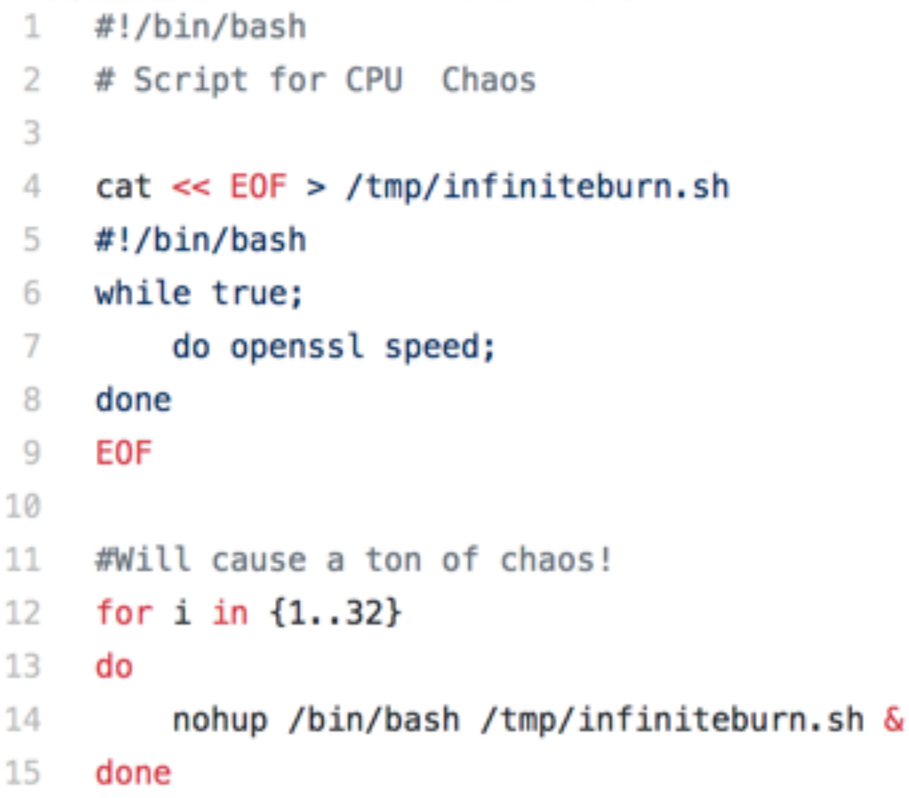
Node 2

159.65.85.169

Node 3

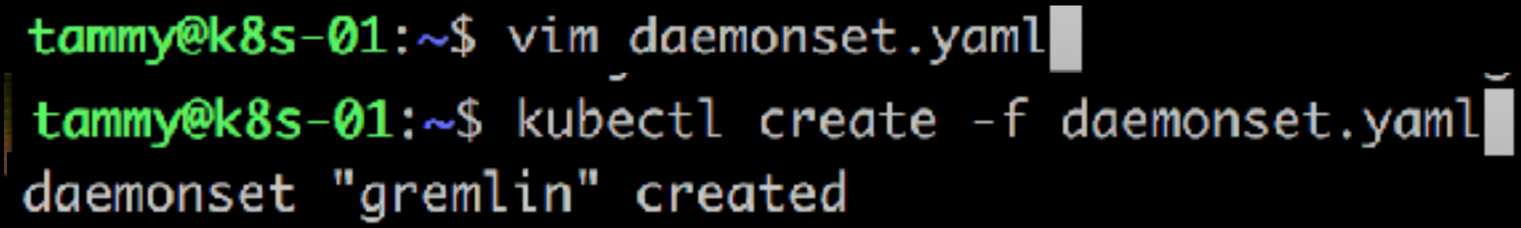
159.65.85.202

Host Level Chaos Engineering With Kubernetes



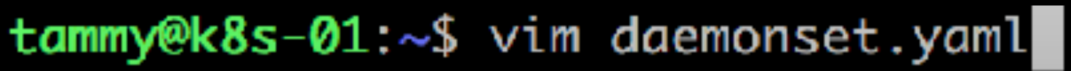
```bash
#!/bin/bash
# Script for CPU  Chaos

cat << EOF > /tmp/infiniteburn.sh
#!/bin/bash
while true;
    do openssl speed;
done
EOF

#Will cause a ton of chaos!
for i in {1..32}
do
    nohup /bin/bash /tmp/infiniteburn.sh &
done
```

Create a Kubernetes Daemonset For Gremlin



```
tammy@k8s-01:~$ vim daemonset.yaml
tammy@k8s-01:~$ kubectl create -f daemonset.yaml
daemonset "gremlin" created
```

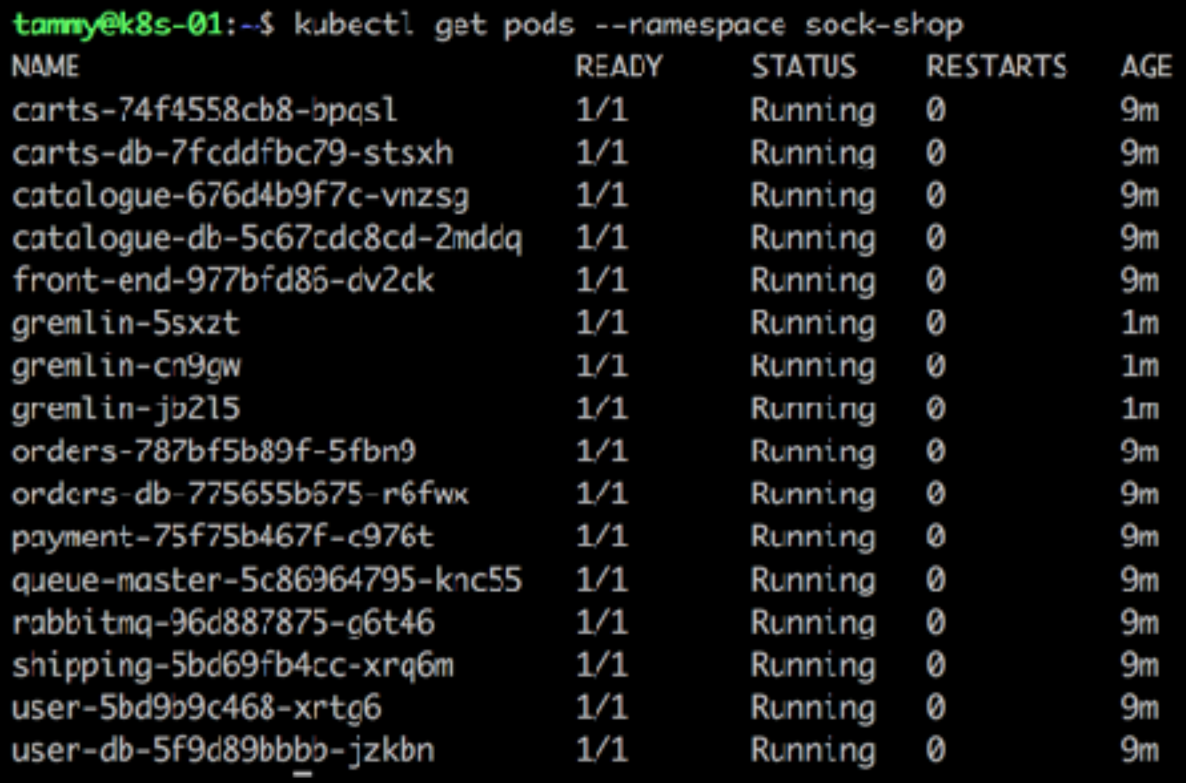
Create a Kubernetes Daemonset For Gremlin



```
tammy@k8s-01:~$ vim daemonset.yaml
```

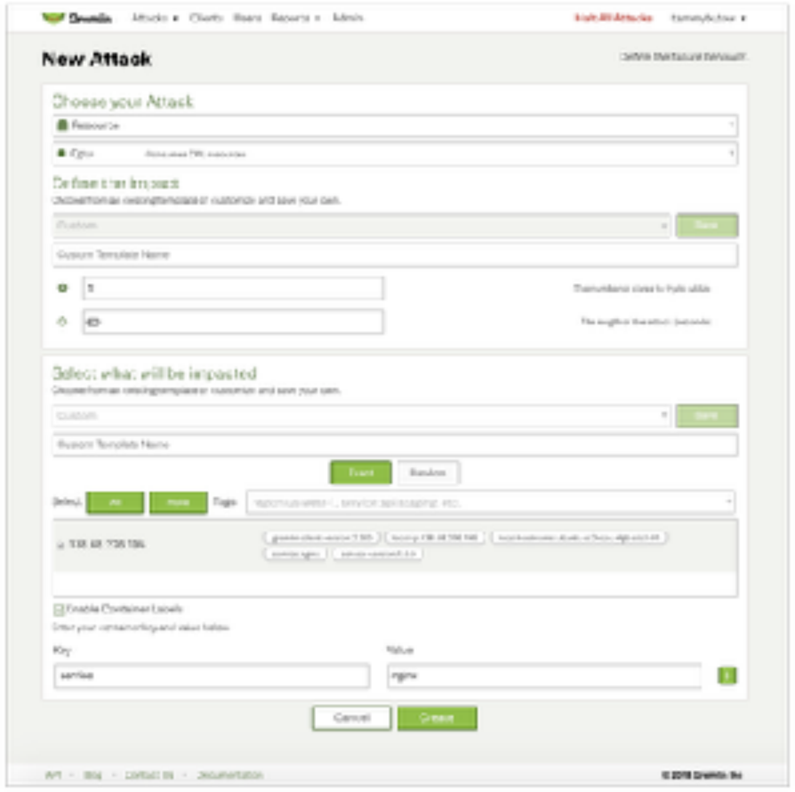
Insert yams

View Your Kubernetes Pods



```
tammy@k8s-01:~$ kubectl get pods --namespace sock-shop
NAME                            READY     STATUS    RESTARTS   AGE
carts-74f4558cb8-bpqsl          1/1       Running   0          9m
carts-db-7fcddfbc79-stsxh       1/1       Running   0          9m
catalogue-676d4b9f7c-vnzsg      1/1       Running   0          9m
catalogue-db-5c67cdc8cd-2mddq   1/1       Running   0          9m
front-end-977bfd86-dv2ck        1/1       Running   0          9m
gremlin-5sxzt                   1/1       Running   0          1m
gremlin-cn9gw                   1/1       Running   0          1m
gremlin-jb215                   1/1       Running   0          1m
orders-787bf5b89f-5fbn9         1/1       Running   0          9m
orders-db-775655b675-r6fwx      1/1       Running   0          9m
payment-75f75b467f-c976t        1/1       Running   0          9m
queue-master-5c86964795-knc55   1/1       Running   0          9m
rabbitmq-96d887875-g6t46        1/1       Running   0          9m
shipping-5bd69fb4cc-xrq6m       1/1       Running   0          9m
user-5bd9b9c468-xrtg6           1/1       Running   0          9m
user-db-5f9d89bbbb-jzkbn        1/1       Running   0          9m
```

Run An Attack From The Gremlin Control Panel



New Attack

Choose your Attack

Resource

Cpu

Define the Impact

Choose from an existing template or customize and save your own.

1 — The number of cores to hang

60 — The length of the attack (seconds)

Select what will be impacted

Choose from an existing template or customize and save your own.

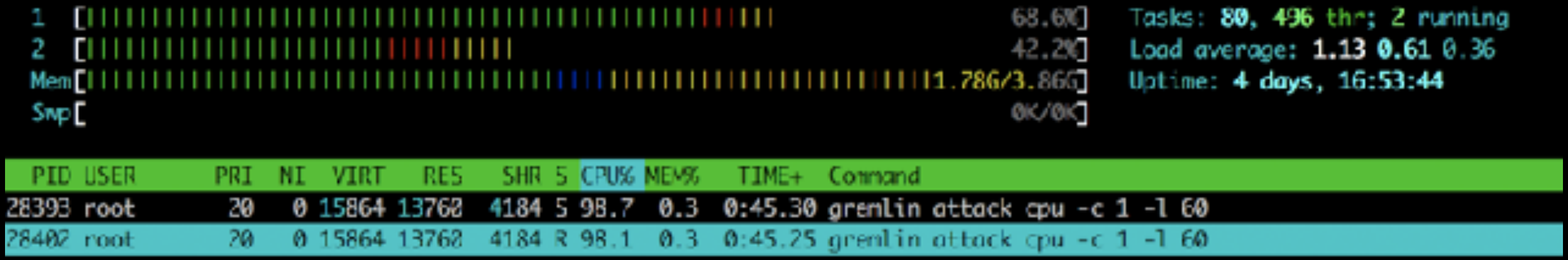
☐ Enable Container Labels

Enter your container key and value below.

Key: service

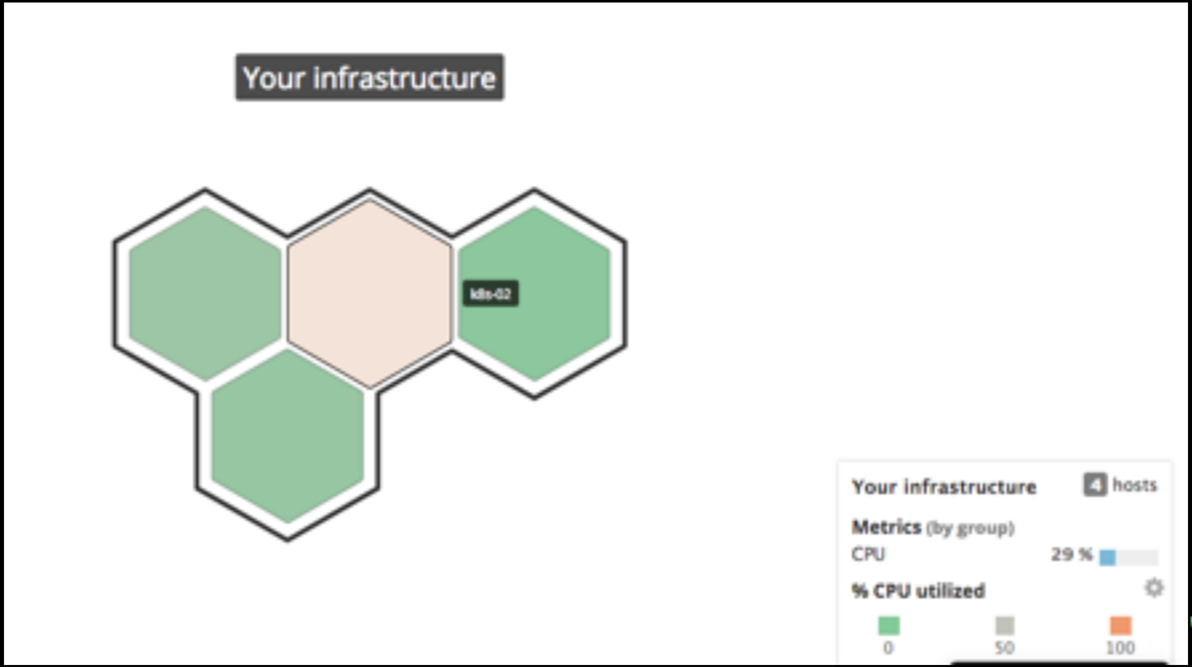
Value: nginx

Monitor Your Chaos Engineering Attack

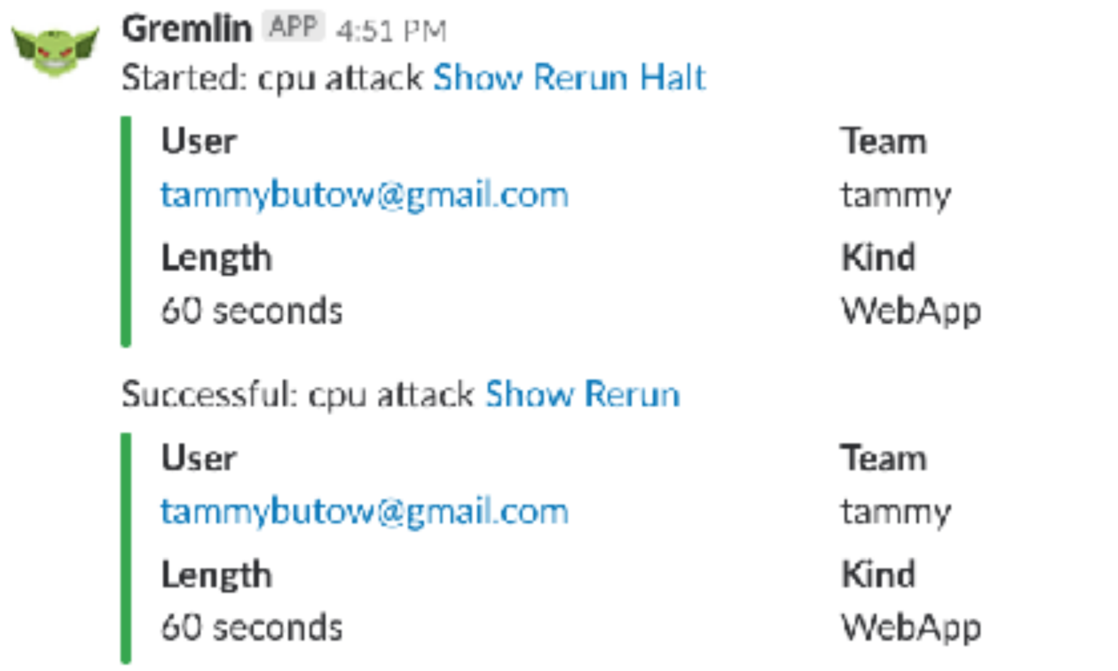


```
  1  [|||||||||||||||||||||||||||||||||||||||||||||||||||||                  68.6%]   Tasks: 80, 496 thr; 2 running
  2  [|||||||||||||||||||||||||||||||||                                      42.2%]   Load average: 1.13 0.61 0.36
  Mem[|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||1.78G/3.86G]   Uptime: 4 days, 16:53:44
  Swp[                                                                       0K/0K]

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
28393 root       20   0 15864 13768  4184 S 98.7  0.3  0:45.30 gremlin attack cpu -c 1 -l 60
28402 root       20   0 15864 13768  4184 R 98.1  0.3  0:45.25 gremlin attack cpu -c 1 -l 60
```

Monitor Your Chaos Engineering Attack



Notify Your Team



Gremlin APP 4:51 PM

Started: cpu attack Show Rerun Halt

User	Team
tammybutow@gmail.com	tammy
Length	Kind
60 seconds	WebApp

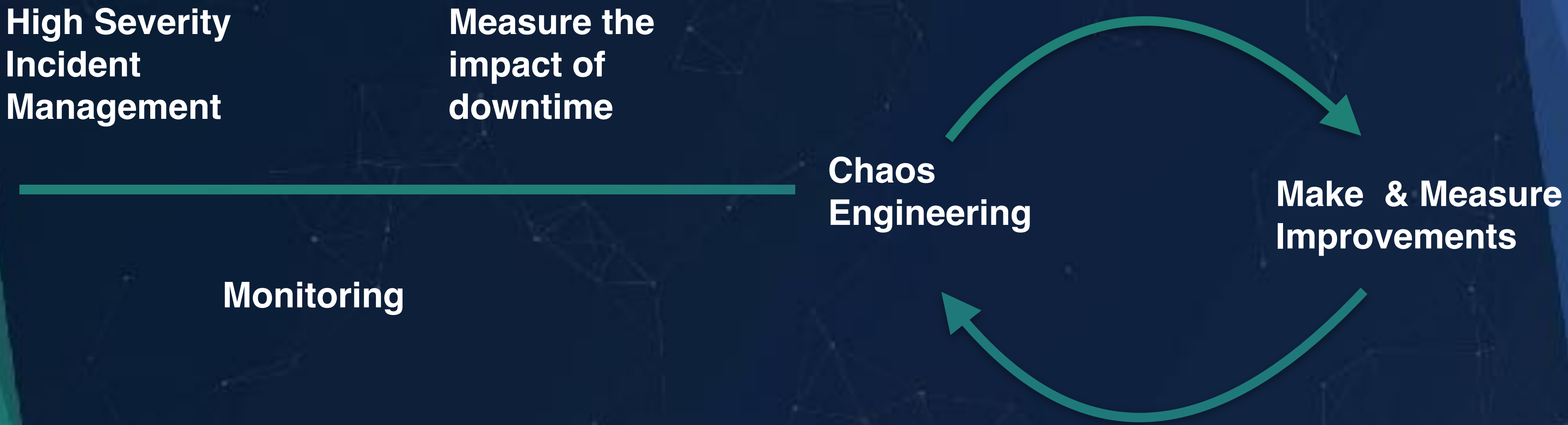
Successful: cpu attack Show Rerun

User	Team
tammybutow@gmail.com	tammy
Length	Kind
60 seconds	WebApp

Let’s Review:

The Path To Chaos Engineering

The Path To Chaos Engineering



High Severity Incident Management

Measure the impact of downtime

Monitoring

Chaos Engineering

Make & Measure Improvements

Blast Radius and Advanced Chaos

High Severity Incident Management

Measure the impact of downtime

Monitoring

Chaos Engineering

Make & Measure Improvements

How do you

Make Improvements?

How do you make improvements?

1. Build - Build a new system / improve existing
2. Borrow - Use open source / contribute to OS
3. Buy - Use 3rd party systems
4. Brush up - GameDays / Team training
5. Break - Chaos Engineering / Failure injection
6. Begone - Decommission systems / delete code

Always Measure Improvements

Tell a story of before and after with metrics

The world needs:

More Resilient Systems

You can create:

More Resilient Systems!

Join us on this journey!

gremlin.com/community

gremlin.com/slack

Thanks!

@tammybutow

gremlin.com