

How to Build Reliable Systems Under Unpredictable Conditions

Smart Approaches in Chaos Engineering, Observability, and Incident Management



Speaker



Benjamin Wilms

CO-FOUNDER & CEO



Setting the scene

What is the **mission** of software development?

Continuously improve & deliver a software solution that delivers value to its users.

**Customers trust a system when it's
consistently good in quality and
performance**

Why is it so hard?

**The complexity of today's systems is
massive**

The complexity of today's **systems is
massive**

Definition: System

Definition: System

Hardware

Software

People

Process

Organisation

Pipeline

...and more.

Improve the Reliability of your System

By unleashing the Power of Chaos Engineering

“Chaos Engineering is thoughtful, planned experiments that reveal the weakness in our sociotechnical systems before they appear in production.”

RUSS MILES

AUTHOR OF “LEARNING CHAOS ENGINEERING”

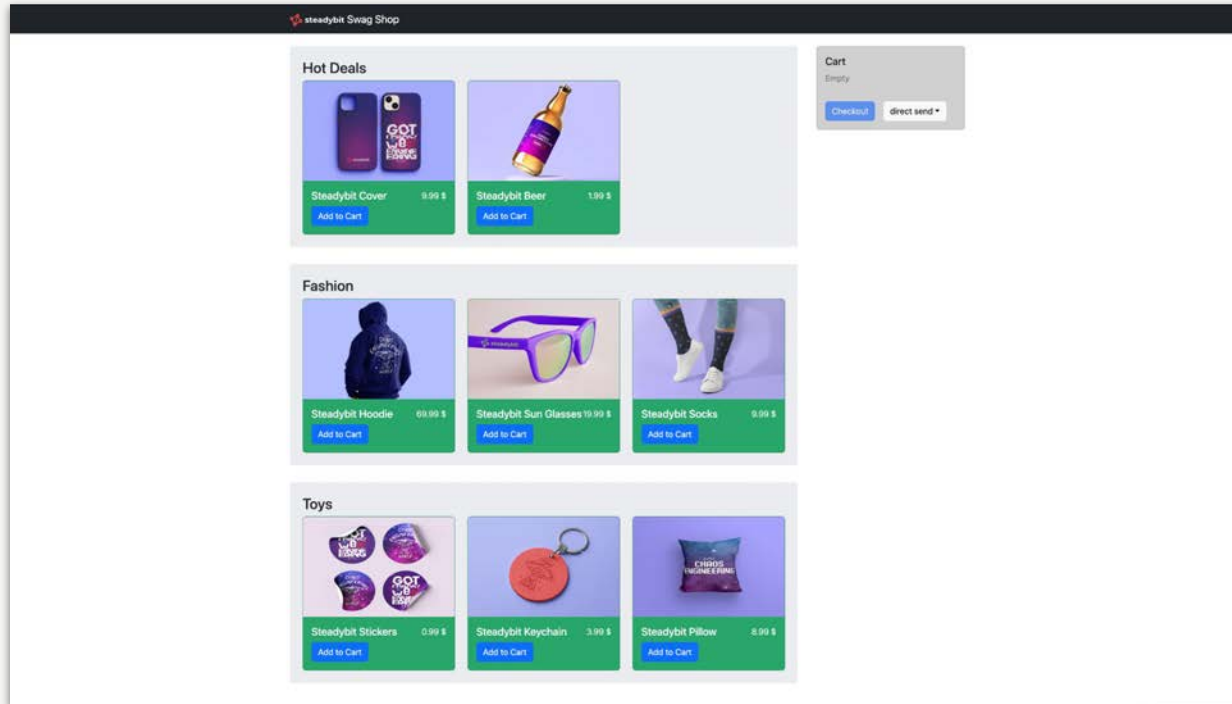
*“Chaos Engineering is thoughtful,
planned experiments that reveal the
weakness in our **sociotechnical systems**
before they appear in production.”*

RUSS MILES

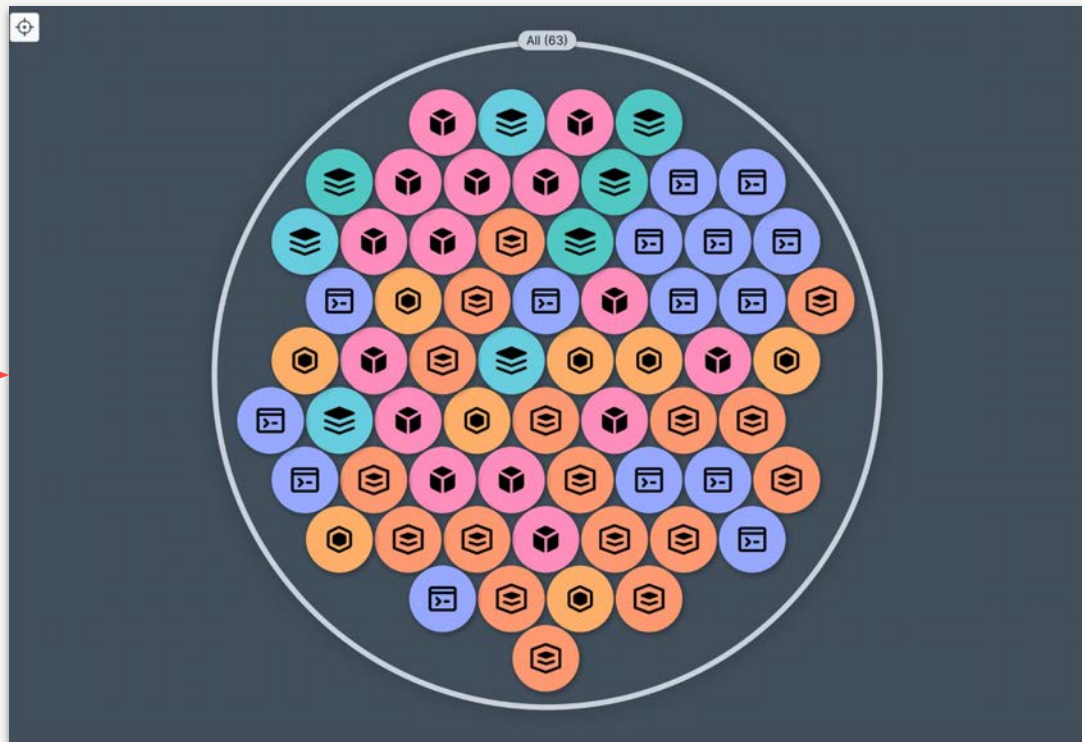
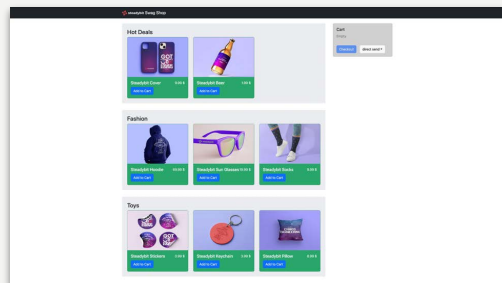
AUTHOR OF “LEARNING CHAOS ENGINEERING”

**Exemplary result a sociotechnical
system creates**

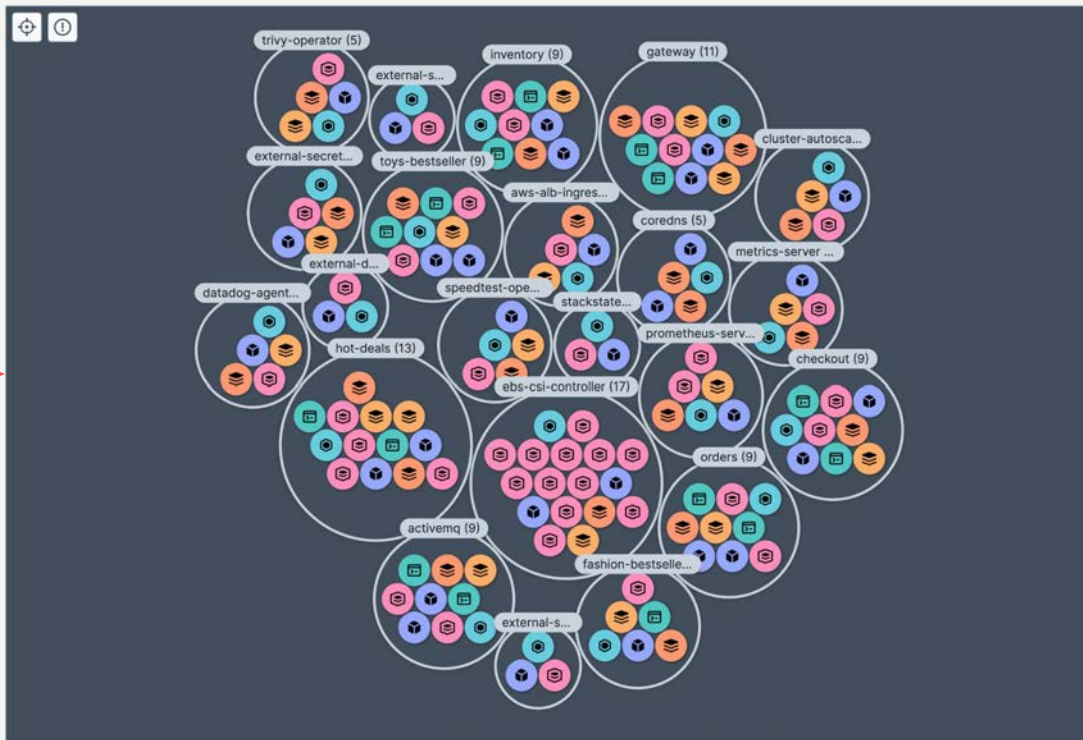
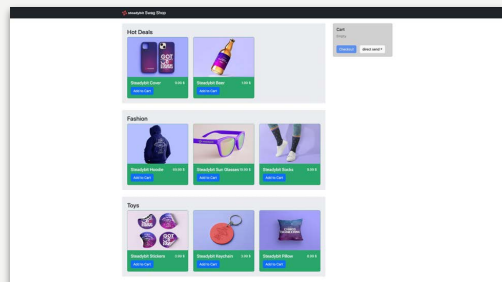
Online shopping system



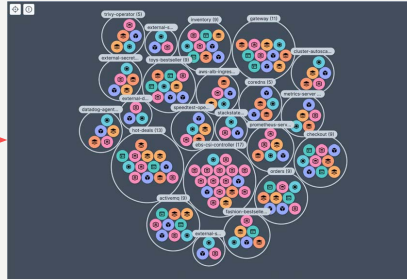
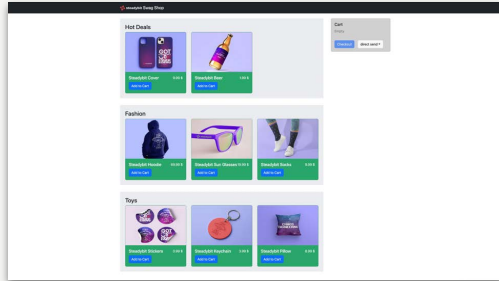
Technical view



Technical view



Organizational view



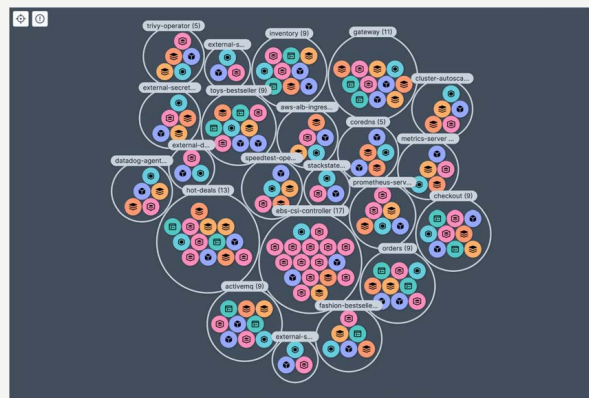
Team A

Team B

Team C

Team D

Organizational view



Team A



Team B



Team C



Team D



Building Resilient Applications Part 1

Increase Systemic Resilience...



...with Steadybit.

Uncover Risk

Identify and Address System Vulnerabilities

Advice analyzes system configurations to identify vulnerabilities and areas for improvement. It provides precise recommendations to enhance system robustness, including optimizing Kubernetes setups and ensuring effective redundancy.



Understand Impact

Verify your Resilience Strategy

Conduct targeted Chaos Engineering experiments to verify that the technical system withstands against potential disruptions.

The screenshot displays the Steadybit Chaos Engineering interface. On the left, a list of experiments is shown, including 'Limit Memory Resources', 'Limit CPU Resources', 'Schedule Pods across AWS Zones' (highlighted), 'Limit Ephemeral Storage Resources', and 'Probes Configured'. A modal for 'Redundant Pod Deployment' is open over the 'Schedule Pods across AWS Zones' experiment. The main panel shows the configuration for 'Schedule Pods across AWS Zones', which is set to 'Kubernetes' and 'Deployment'. A yellow callout box states: 'VALIDATION NEEDED: Your pods are spread across multiple zones. Now, validate your redundancy by simulating an outage of one zone.' Below this, a table of 'Recommended Validations' shows 'Degraded Performance of Availability Zone' as disabled and 'ADM-123 Availability Zone Outage' as completed on 02/03/2024 at 17:13:00. A 'Create Experiment' button is visible. The bottom of the modal shows the 'Motivation' for the experiment: 'An availability zone can be unavailable as they are not redundantly designed. In order to survive an outage of the availability zone eu-central-1b you should spread your Kubernetes pods across multiple availability zones.'

Filter by technology:

Kubernetes (6)

Filter by:

Limit Memory Resources

Limit CPU Resources

Schedule Pods across AWS Zones

Limit Ephemeral Storage Resources

Probes Configured

Redundant Pod Deployment

Schedule Pods on different Hosts

Schedule Pods across AWS Zones

VALIDATION NEEDED

Your pods are spread across multiple zones. Now, validate your redundancy by simulating an outage of one zone.

Recommended Validations

Validated	Experiment Name
<input type="checkbox"/>	Degraded Performance of Availability Zone
<input checked="" type="checkbox"/>	ADM-123 Availability Zone Outage

Completed 02/03/2024, 17:13:00

Create Experiment

Action Implemented

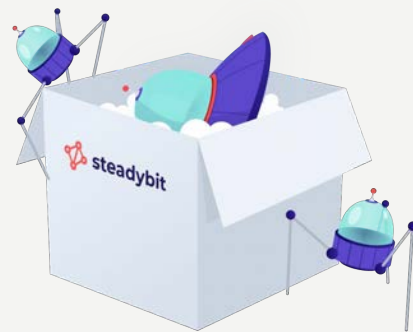
Motivation

An availability zone can be unavailable as they are not redundantly designed. In order to survive an outage of the availability zone eu-central-1b you should spread your Kubernetes pods across multiple availability zones.

Uncover Risk & Understand Impact

Demo Landscape & Reliability Advice

<https://platform.steadybit.com/landscape/explore/997ca4ba-ea7a-4d9d-98aa-de12362bc73f?tenant=demo~>



chrome

FileEditViewHistoryBookmarksProfilesTabWindowHelp

Building Resilient Application xSteadybit / Experiments x

platform.steadybit.com/experiments/edit/SHOP-981/executions;page=0-/54001/attack?tenant=demo-

Experiments / SHOP-981 Single Pod Failure of fashion-bestseller

DesignRuns

Recent Runs

#54001 Global 14/05/2024, 14:25:30

#54000 Global 14/05/2024, 14:22:51

#54001

Created14/05/2024, 14:25:30 by Benjamin Wilms

Attack MonitorAgent LogTracing

1

TODO VALIDATION: INVARIANT: fashion-bestseller's features work within expected success rates

2

GIVEN: All pods are ready

THEN: All pods become ready again within 60s

3

THEN: One pod is detected failing

4

Show Kubernetes events from the cluster

5

Show Pod Count Metrics for the cluster

0s15s30s45s60s75s90s105s120s135s150s

Run Status

14:25:30 Establishing connection to agents...

14:25:30 Experiment run created by Benjamin Wilms.

Deployment Readiness

Kubernetes Events

>>> FAST FWD

Building Resilient Applications

Part 2

Increase Systemic Resilience...

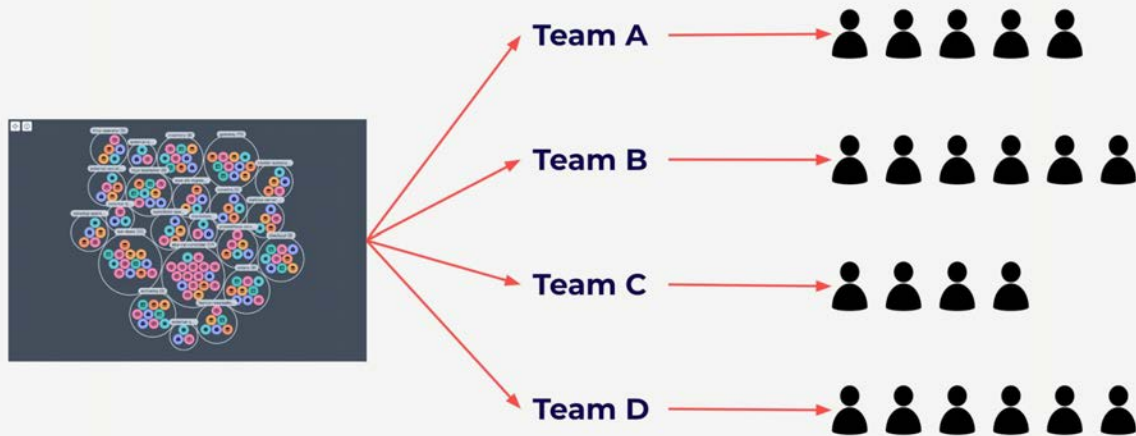


...with Steadybit.

Strengthen Organisation

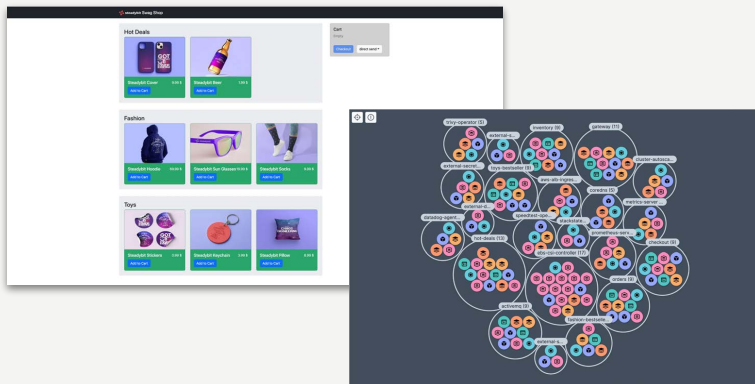
Verify interaction of all necessary players

After continuously improving and reviewing our technical system, we turn our attention to the sociotechnical system

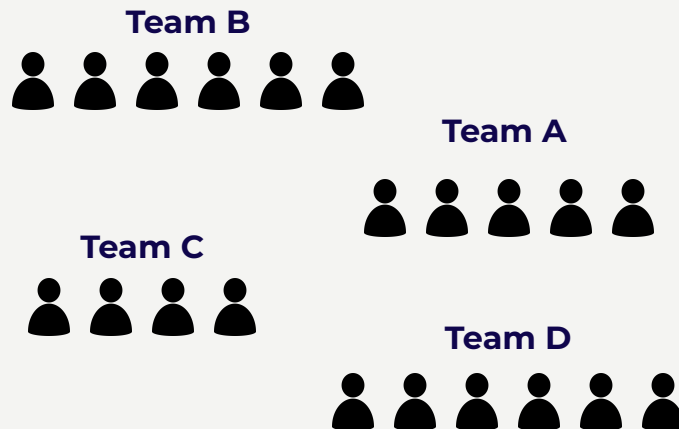


Strengthen Organisation

Technical System



Organisational System



Strengthen Organisation

Scenario

Latency spikes in backend service *Hot Deals* followed by total outage in the central *Gateway* service.

Hypothesis

Our monitoring recognises this scenario within 90s and reports it to the relevant team on call via PagerDuty. The incident is opened and acknowledged within 3 minutes.

The system normalises within 3 minutes and the incident and our on-call team determines that the resilient system has recovered and is working normally. The incident is resolved within 4 minutes..

Strengthen Organisation

Interaction

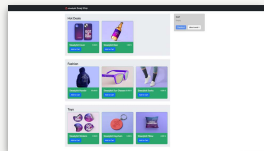
PagerDuty



Strengthen Organisation

Timeline

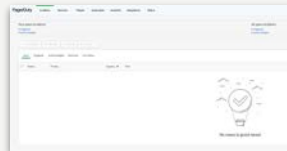
Latency spike
followed by total
outage



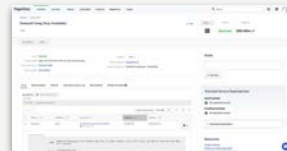
Monitoring
Event
within 90s



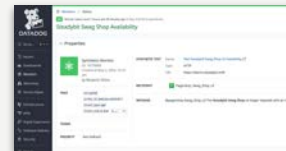
Incident
Triggered in
PagerDuty
within 100s



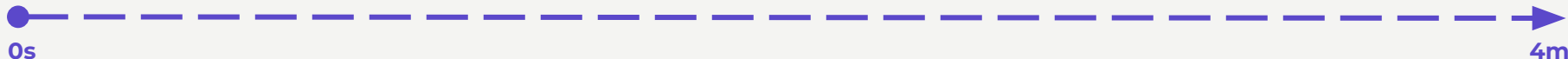
Incident
Acknowledged
in PagerDuty
within 3m



System
OK
after 3m



Incident
Resolved in
PagerDuty
within 4m



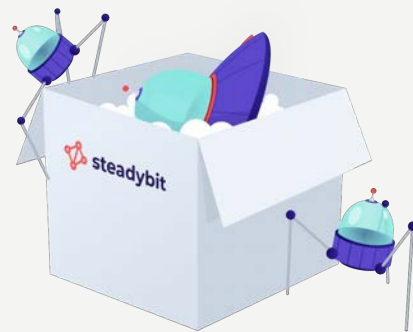
It's **not** about testing how fast
an individual person fixes an
outage.

It's about how good the organisation is at detecting and fixing faults and how processes are coordinated.

Strengthen Organisation

Demo scenario design in Steadybit

<https://platform.steadybit.com/experiments/edit/SHOP-972/design?tenant=demo~&team=SHOP~>



Chrome

File Edit View History Bookmarks Profiles Tab Window Help

Steadybit / Experiments

platform.steadybit.com/experiments/edit/SHOP-972/executions;page=0-J53580/attack?tenant=demo-

New Chrome available

Experiments / SHOP-972 Webinar Showcase

DesignRuns

00:04:59

Recent Runs

#53580 Global03/05/2024, 17:18:19

#53580Running03/05/2024, 17:18:19 by Benjamin Wilms

Attack MonitorAgent LogTracing

17:18:20

1Datadog Monitor Status = OK | Shopping UI Availability

2Default system behaviourWait for Monitoring Status "Alert"Datadog Monitor Status = ALERT | Shopping UI Availability

3Default system behaviourWait for Monitoring Status "Ok"Datadog Monitor Status = OK | Shopping UI Availability

4Default system behaviourWait for Incident "Triggered" in PagerDutyCheck PagerDuty - Incident Trigg

5Default system behaviourWait for Incident "Acknowledged" in PagerDutyCheck PagerDuty - Incident Ack

6Default system behaviourWait for Incident "Resolved" in PagerDutyCheck PagerDuty - Incident Res

7Default system behaviourLatency spikeTotal outage GATEWAY

Run Status

17:18:25Datadog Monitor Status = OK | Shopping UI Availability

17:18:25Default system behaviour

17:18:25Default system behaviour

17:18:25Default system behaviour

17:18:25Default system behaviour

HTTP Responses for https://api.pagerduty.com/incidents?urgency=high&statuses[]=triggered&service_ids[]=P3V6VIA&sort_by=created_at_DESC

Waiting for metrics...

HTTP Responses for https://api.pagerduty.com/incidents?urgency=high&statuses[]=acknowledged&service_ids[]=P3V6VIA&sort_by=created_at_DESC

Waiting for metrics...

HTTP Responses for https://api.pagerduty.com/incidents?urgency=high&statuses[]=resolved&service_ids[]=P3V6VIA&sort_by=created_at_DESC

Waiting for metrics...

Datadog Monitor Status

PRE SYSTEM CHECK

Waiting for metrics...

Recap

Contact details



Benjamin Wilms

steadybit.com

signup.steadybit.com

**Thank you
for your
time.**

