Cloud Operations at Scale

Managing 250 Kubernetes Clusters at Zalando

Heinrich Hartmann & Mikkel Larsen

Zalando SE

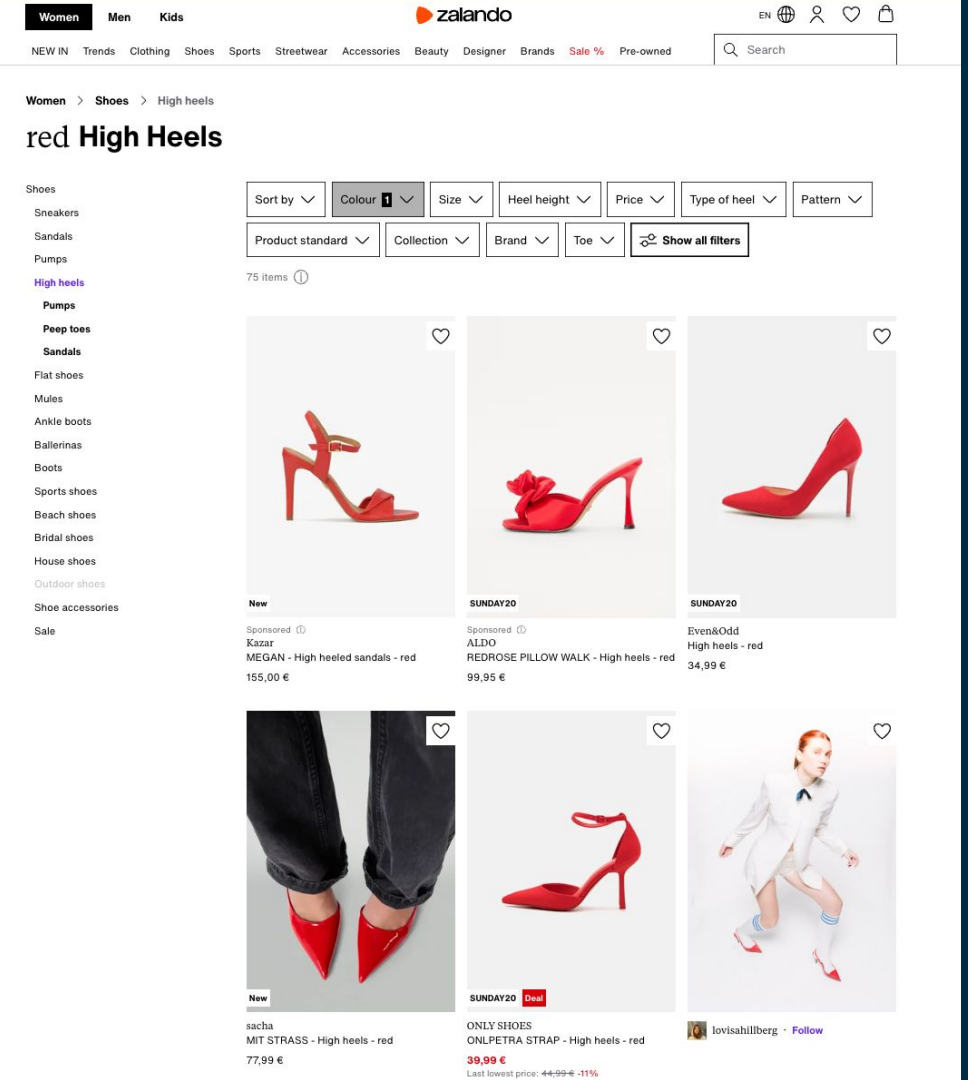
About us

Heinrich Hartmann

- Senior Principal SRE at Zalando
- 10 years of Reliability Engineering
- Led SRE Department for 2.5 years
- Chief Data Scientist @ Circonus
- PhD in Mathematics

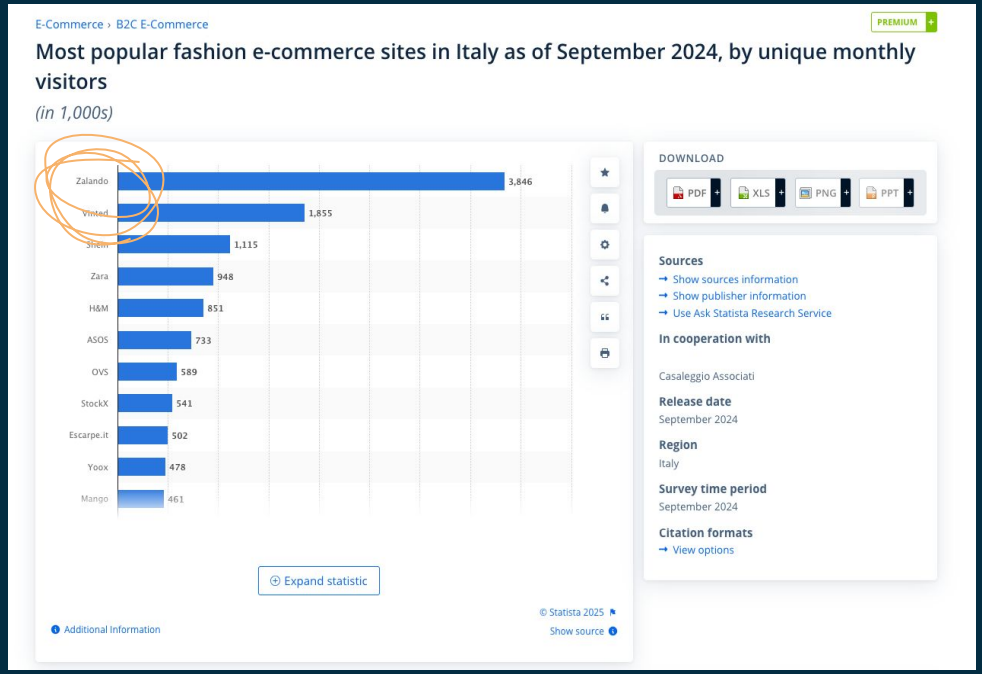
Mikkel Larsen

- Principal Engineer at Zalando
- 9 years of Cloud Infra (AWS/Kubernetes)
- Open Source Maintainer



- Leading Fashion Platform in Europe
- Founded 2008
- 50M+ Customers
- 10B+ Revenue
- 3K+ Software Engineers

Grazie mille!



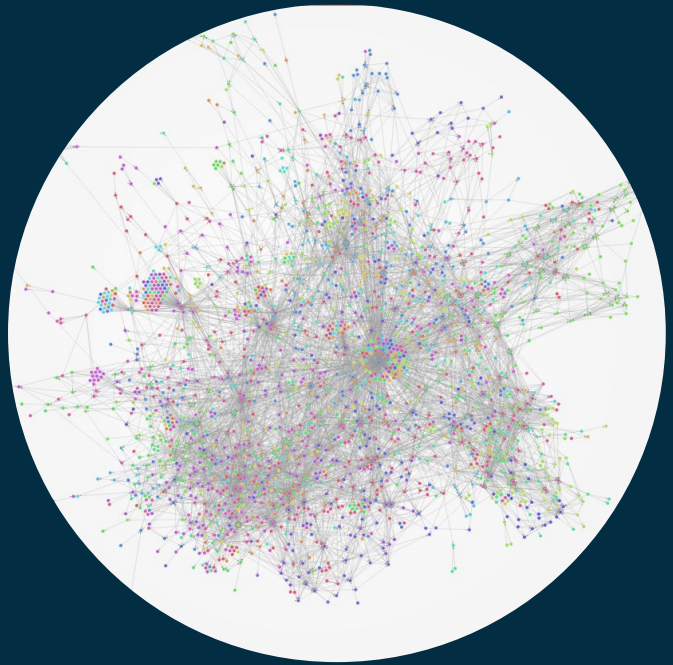
source: https://www.statista.com/statistics/1265105/most-popular-fashion-e-commerce-sites-italy/

Agenda

1. 🔭 Cloud Strategy
2. 🚀 Cloud Operations
3. 💥 Cloud Incidents

Cloud Strategy

Zalando Scale



Zalando Service Graph (2019)

Business

- 200K+ HTTP requests per second
- Thousands of orders per minute peak

Technology

- 3000+ Microservices
- 100K+ Containers
- 10K+ Nodes

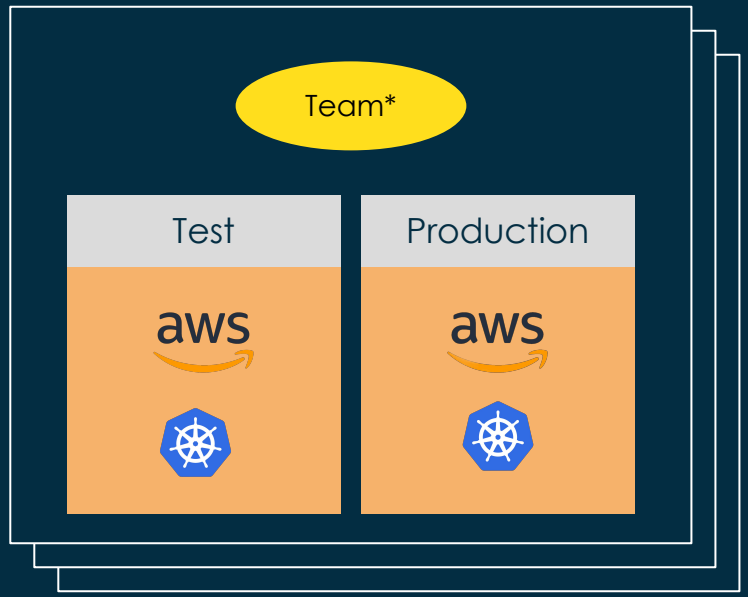
source: https://corporate.zalando.com/en

Zalando Cloud Strategy

- No Data Centers*
- Single Cloud (AWS)*
- Single Region (Frankfurt)*
- Everything in Kubernetes*
- Leverage AWS Managed Services**

*) almost **) responsibly

One Team – One Account – One Cluster



300 AWS Accounts, 250 K8S Clusters

Benefits

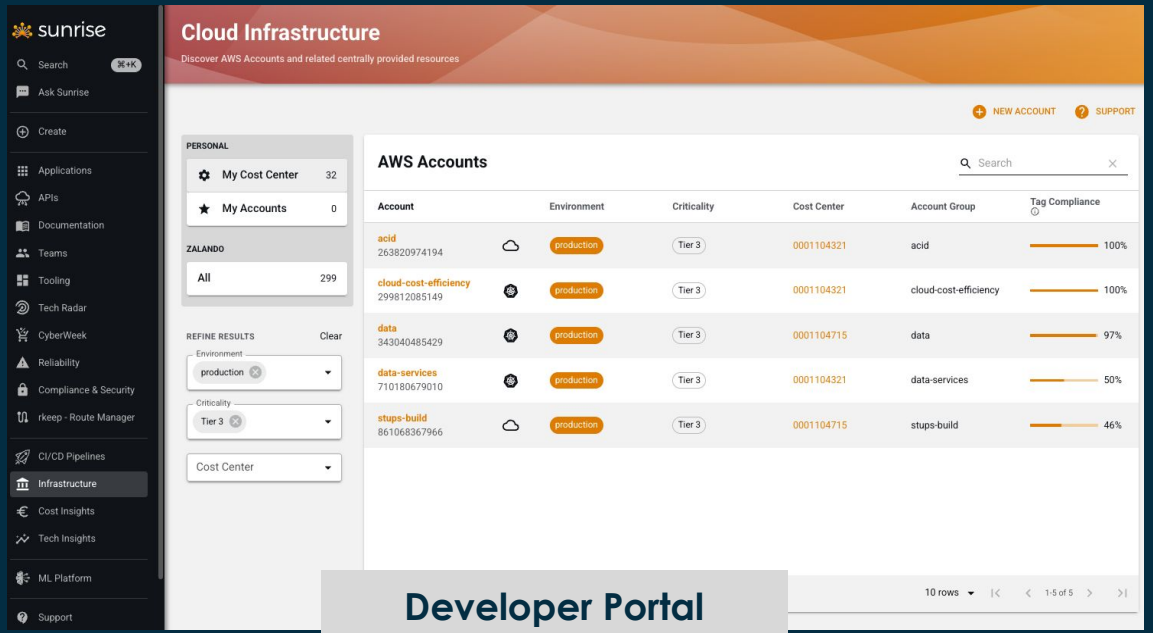
- Team Autonomy (Conway's Law)
- Reliability & Security
- Cost attribution
- Access management
- Avoid K8S scaling limits (2K nodes)

Drawbacks

- Higher Cost
- Lower Efficiency
- Requires Automation
- …

*) or group of up to 5 teams in a given domain.

Cloud Platform Services



Developer Portal
(Backstage.io)

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: testapp   # name of the Pod
  labels:
    application: testapp   # name of the application this Pod belongs to
    component: backend     # name of the component of the application
spec:
  containers:
  # our Pod has just one container
  - name: testapp   # name of the container
    image: container-registry-test.zalando.net/teapot/training-example:master-44
    ports:
```

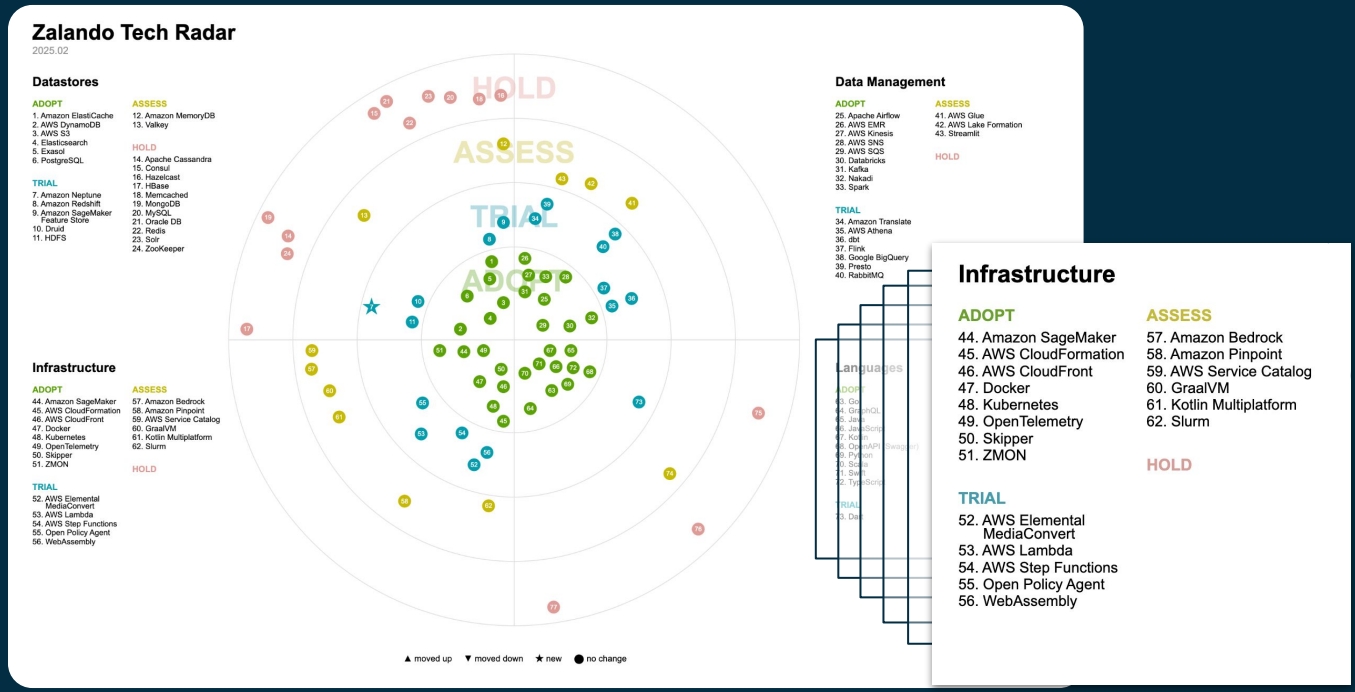
***.yaml on GitHub**

```
~
❯ pipx install zalando-cli-bundle --include-deps
  installed package zalando-cli-bundle 24.17.0, installed using Python 3.12.8
  These apps are now globally available
    - docker-credential-pierone
    - jp.py
    - kio
    - markdown-it
    - natsort
    - normalizer
    - pierone
    - piu
    - pygmentize
    - zalando-cli-bundle
    - zaws
    - zign
    - zkubectl
    - ztoken
done! ✨ 🌟 ✨
```

CLI Tools

Cloud Governance w/ TechRadar

https://opensource.zalando.com/tech-radar/



Infrastructure

ADOPT

44. Amazon SageMaker
45. AWS CloudFormation
46. AWS CloudFront
47. Docker
48. Kubernetes
49. OpenTelemetry
50. Skipper
51. ZMON

ASSESS

57. Amazon Bedrock
58. Amazon Pinpoint
59. AWS Service Catalog
60. GraalVM
61. Kotlin Multiplatform
62. Slurm

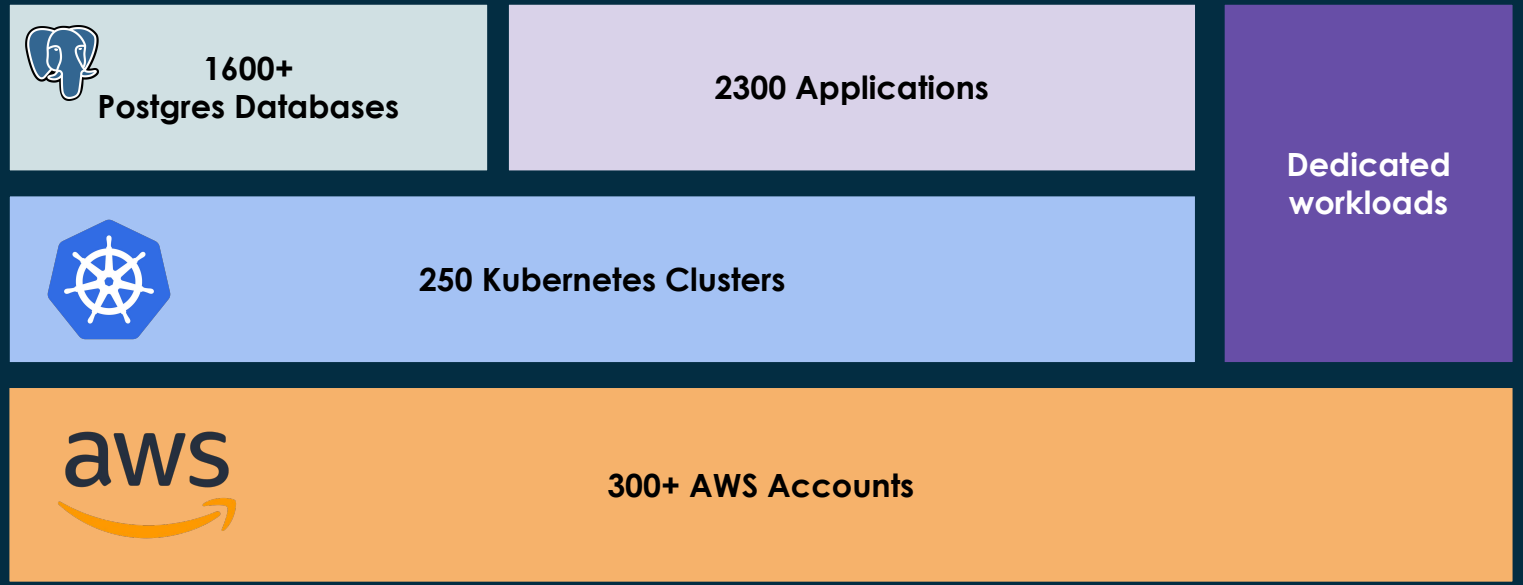
HOLD

TRIAL

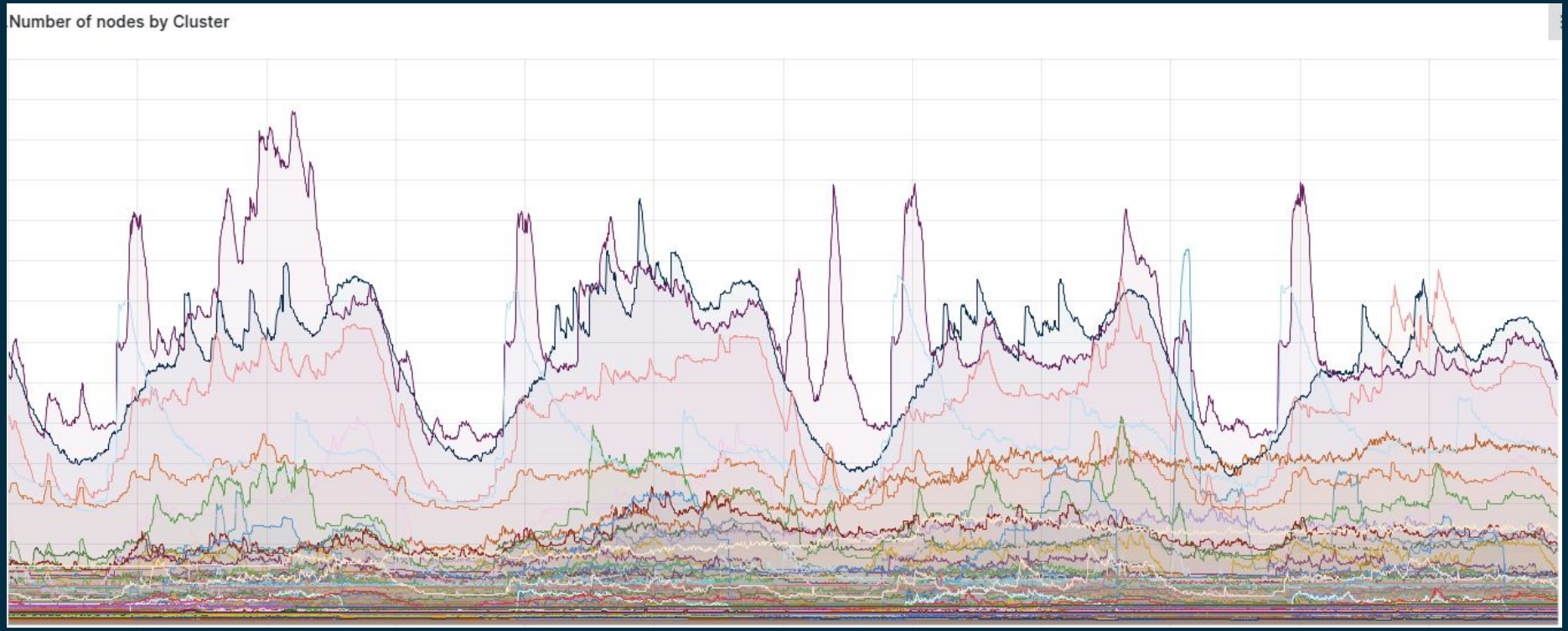
52. AWS Elemental MediaConvert
53. AWS Lambda
54. AWS Step Functions
55. Open Policy Agent
56. WebAssembly

Cloud Operations

Cloud Operations - Scale



Cloud Operations - Dynamic Scale



Cloud Operations - Original “Philosophy”

- **No pet Infrastructure**

  Cluster and Account configuration is managed as code in central repositories.

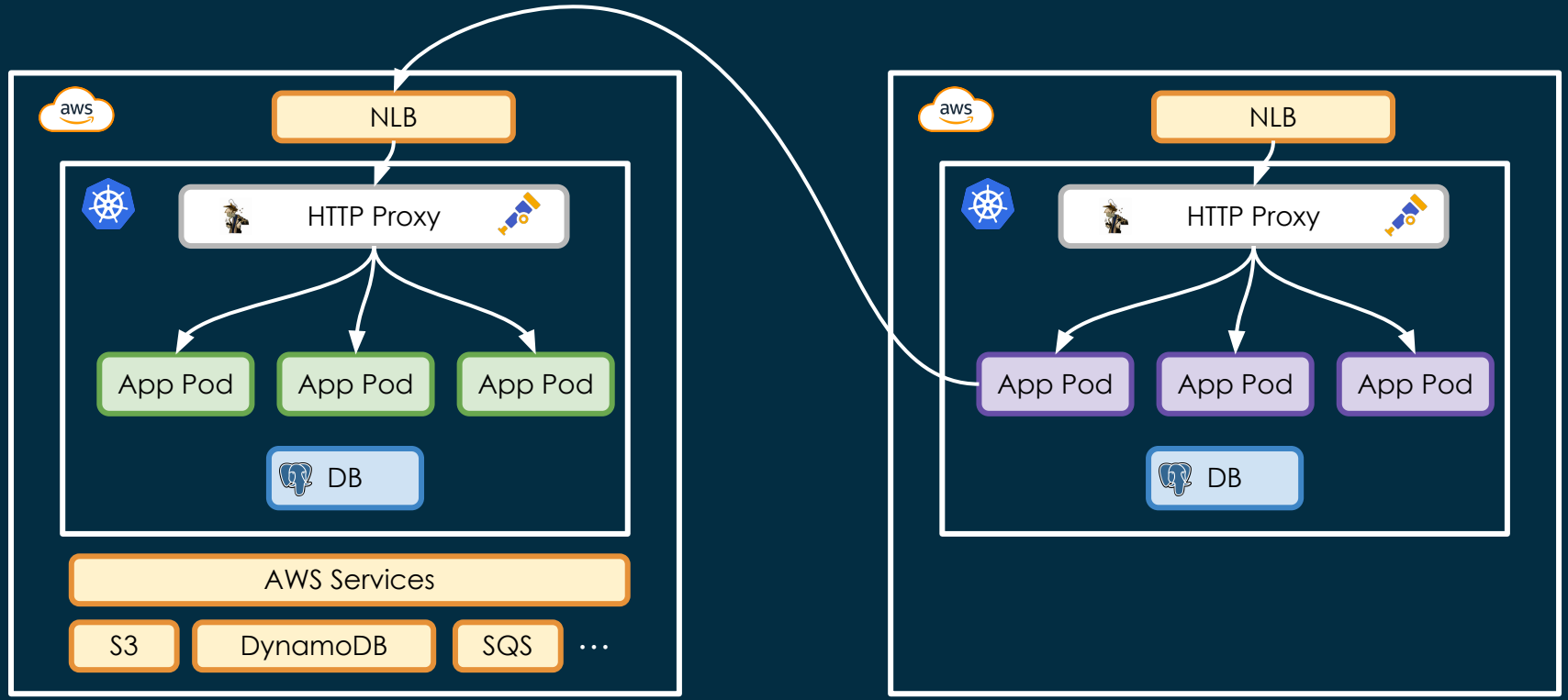
- **Always provide the latest stable Kubernetes version**

  Latest features and quick to address possible vulnerabilities.

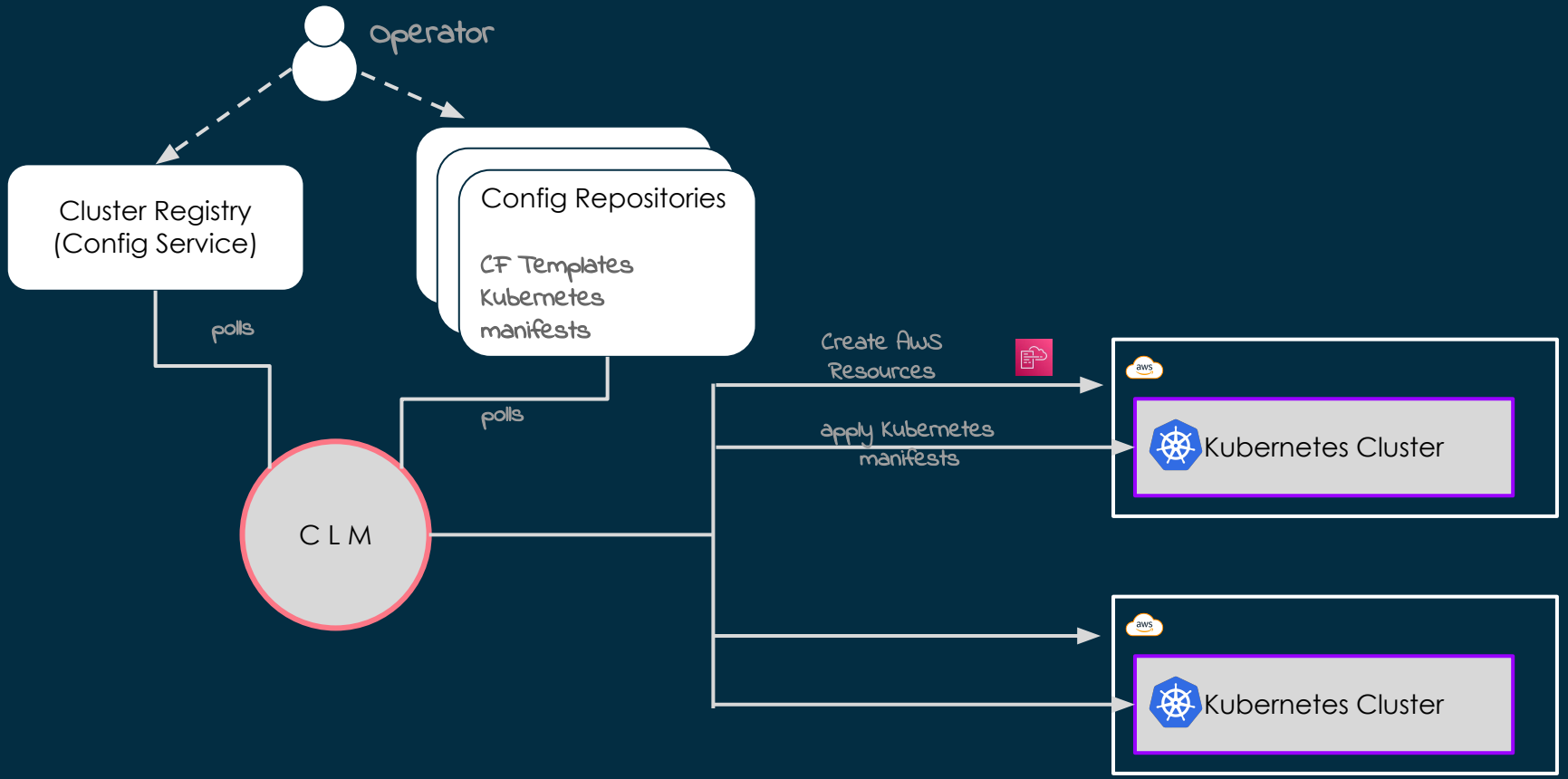
- **Continuous and non-disruptive cluster updates**

  No maintenance windows.

Cloud Operations - Architecture



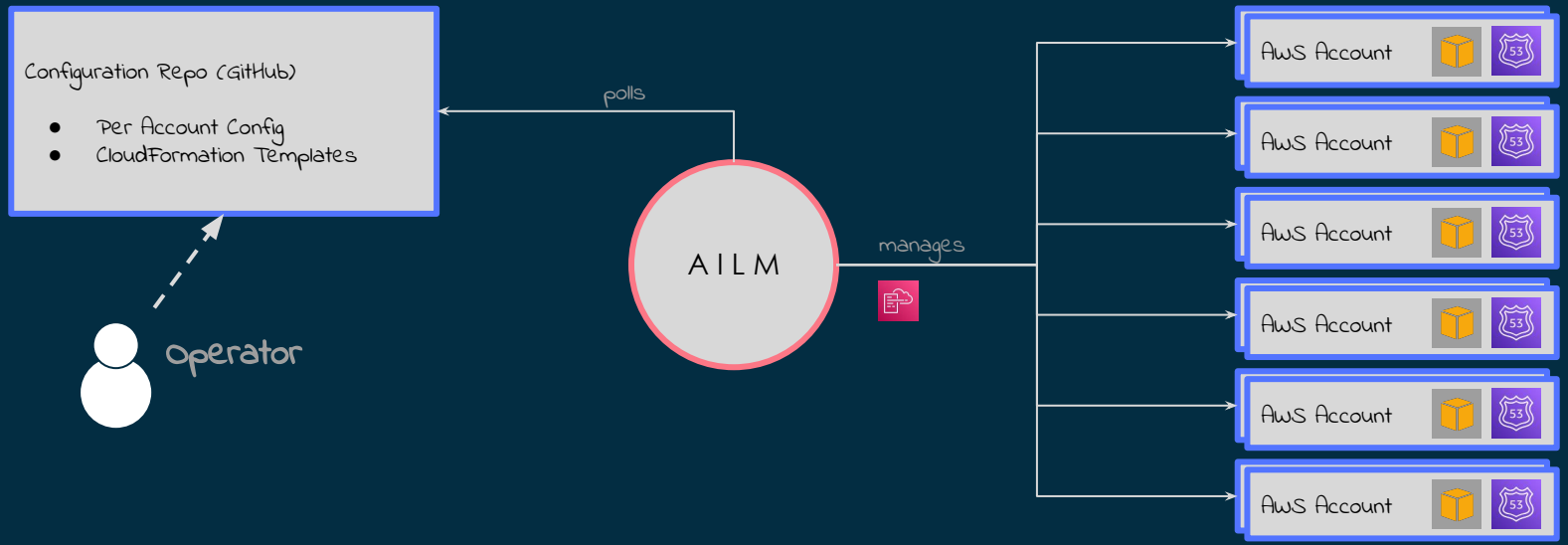
Cloud Operations - Infrastructure as Code (Kubernetes)



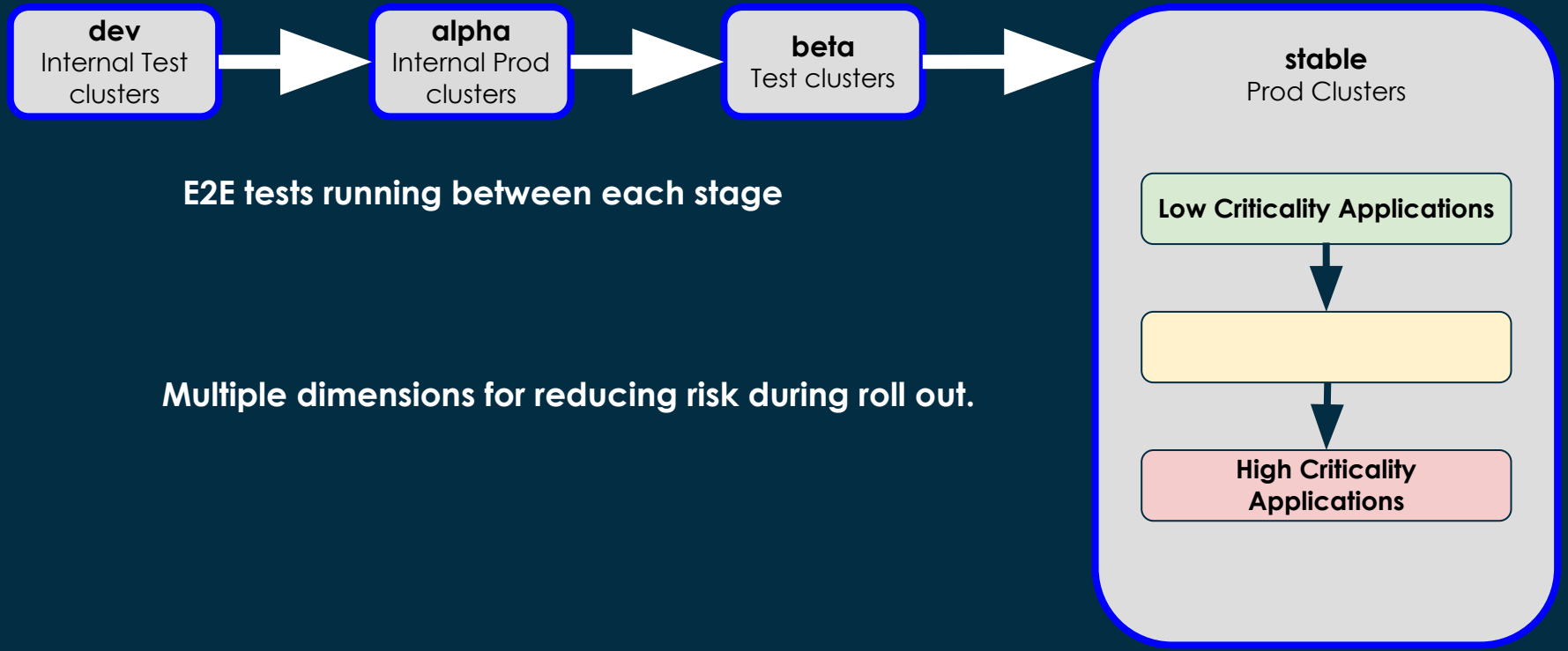
Cloud Operations - Infrastructure as Code (AWS)

Manages all Base Infrastructure

- VPC
- Hosted Zones
- Certificates
- IAM Policies



Cloud Operations - Staged roll out (Kubernetes)

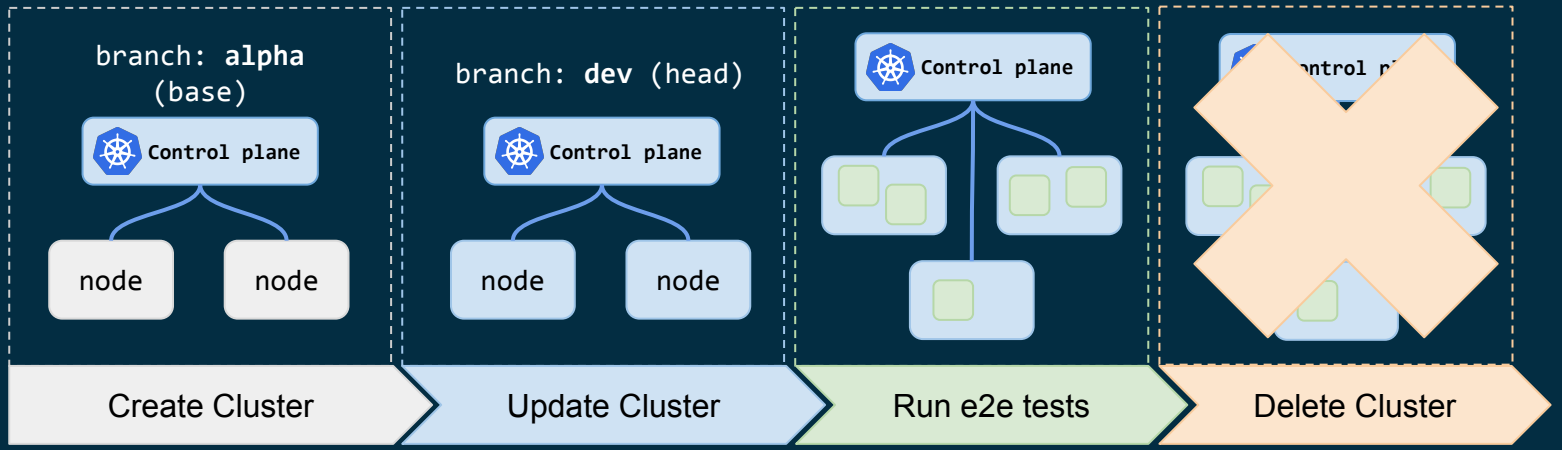


E2E tests running between each stage

Multiple dimensions for reducing risk during roll out.

Cloud Operations - e2e test on every change

Testing **dev** to **alpha** upgrade



🕒 **1 hour for a full e2e test run**

Cloud Operations - “Philosophy” vs. Practice

- No pet Infrastructure
  - Almost everything in code.
  - Many per account/cluster configurations.
- Always provide the latest stable Kubernetes version
  - Oldest clusters were upgraded from Kubernetes v1.4 through to v1.31. (8.5 years)
- Continuous and non-disruptive cluster updates
  - Limiting monthly node rotations to reduce disruption of stateful workloads.

Cloud Incidents

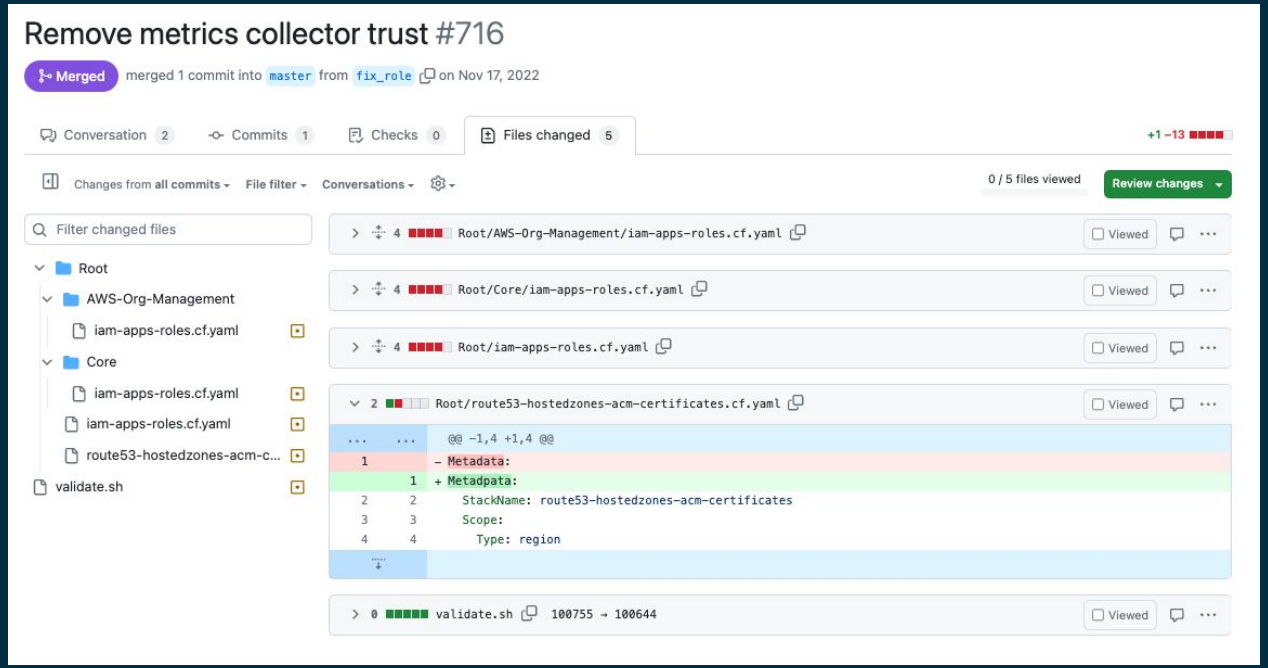
The Setting

It's Tuesday, November 17th 2022

...

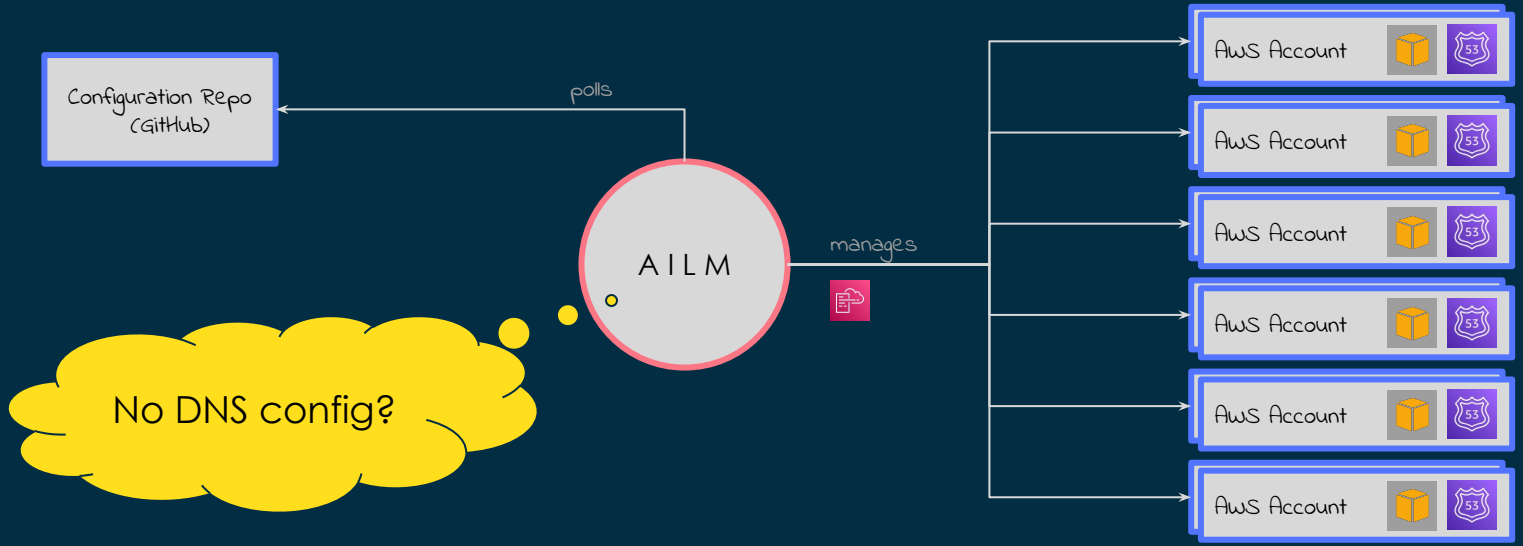
The Trigger

12:56 - 🚢 PR #716 gets merged



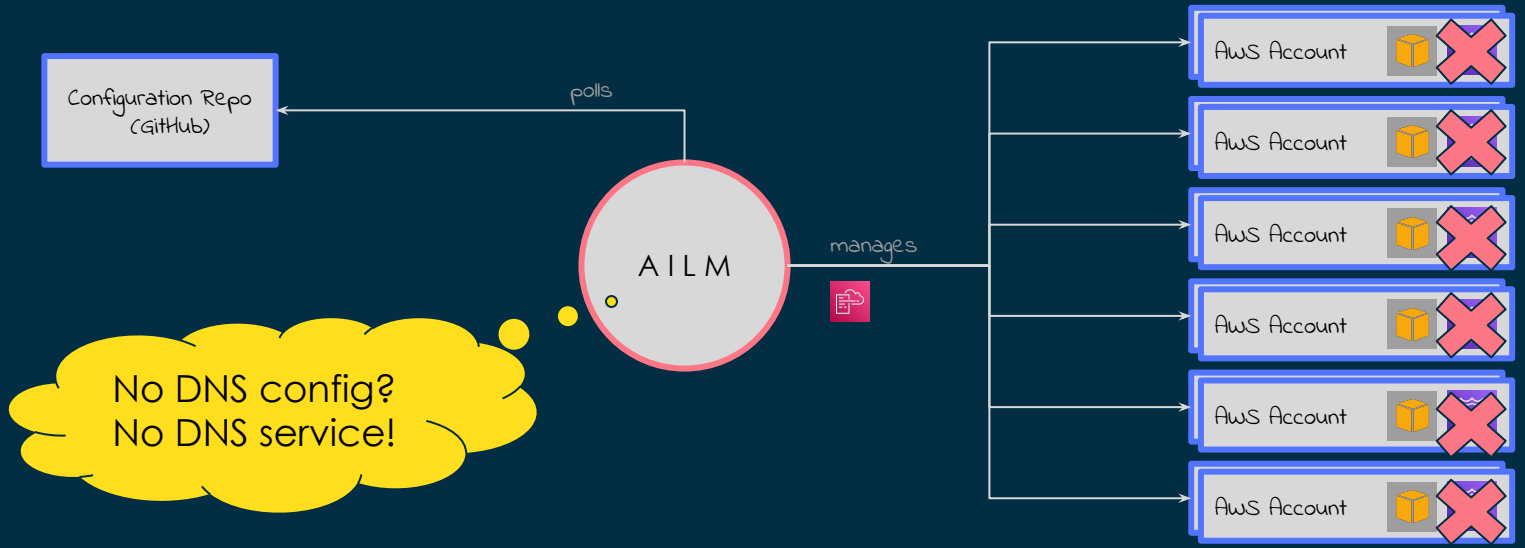
Automation takes over

12:58 - 🤖 AWS Lifecycle Manager processes change



Automation takes over

12:58 - 🤖 AWS Lifecycle Manager processes change



The Fallout

13:10 - 🔥 Zalando goes down

The Fallout

13:10 - 🔥 Monitoring goes also down

The Cleanup

- 13:23 - PR #716 is identified as culprit
- 13:38 - PR #716 is reverted - without success
- 13:40 - 👨‍💻 Manual restoring of DNS entries starts

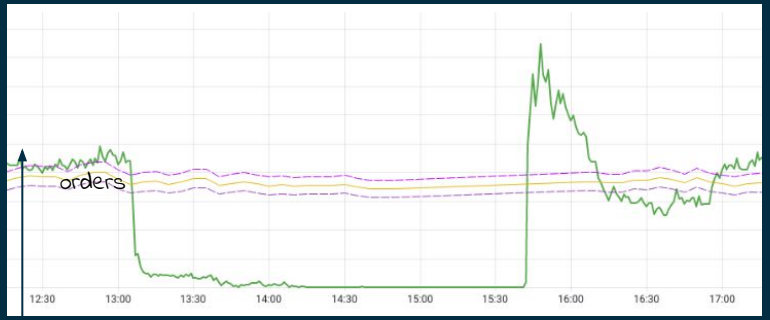
  ~ 200 Engineers in Incident Video Meeting

aws-chooser.zalando.net	18.156.81.28	18.198.223.174	52.57.226.107
emergency-access-service.stups.zalan.do	18.156.81.28	18.198.223.174	52.57.226.107
kube-2.stups.zalan.do			
legacy-teams-api.corporate-iam.zalan.do			
kube-web-view.zalando.net			
zmon.zalando.net			
registry.opensource.zalan.do			
pierone.stups.zalan.do			
container-registry.zalando.net			
info.services.auth.zalando.com			
platform-iam-tokeninfo.corporate-iam.zalan.do			
sunrise.platform-infrastructure.zalan.do.			

prio		cname- Status		External DNS fixed?
product-availability	Rodrigo	Done	External DNS Logs	Yes
fulfillment	Katyanna	Done	External DNS Logs	Yes
tx-core	Rodrigo	Done	External DNS Logs	Yes
fashion-store	Katyanna	Done	External DNS Logs	Yes
information-experience	Noor	Done	External DNS Logs	Yes
customer-data-platform	Mahmoud	Done	External DNS Logs	Yes
dcis	Noor	Done	External DNS Logs	Yes
search	Katyanna	Done	External DNS Logs	Yes
inbox	Martin	Done	External DNS Logs	Yes
post-purchase	Martin	Done	External DNS Logs	Yes
checkout	Noor	Done	External DNS Logs	Yes
cart	Zak	Done	External DNS Logs	Yes
db	Zak	Done	External DNS Logs	Yes

Restoring Service

- 14:30 - 🛠️ Monitoring & Ops tools start to recover
- 14:53 - 👠 Catalog starts to recover
- 15:41 - 💵 Order processing start to recover
- 17:00 - 📦 Warehouse services are restored
- 20:02 - ✅ All systems fully recovered



The Learnings

- **Safeguard Infrastructure Changes**

  Validation, Preview, Staged-rollout.

- **Tighten Deployment Policy**

  Applies to configuration (not just code) and tooling that *can* impact production.

- **Segregate Infrastructure for Monitoring & Ops Tooling**

  Separate rollout stages for internal services.

- **Blameless Post Mortem Culture**

  Human error is never the root cause. Fix process to catch mistakes.

Infrastructure Incident Patterns

- **Sharp Edges with Running K8S**
  ⇒ Fail, Fix, Repeat
  ⇒ Upstream code & share learnings https://k8s.af/.

- **Change Failures with "Supertools"**
  ⇒ Validation, Preview & Staged rollout

- **Overload of Kubernetes Backplane**
  ⇒ Overscale. Move to AWS EKS.

- **Upstream AWS Incidents** (Networking Problems, AZ Outages, etc.)

TL;DR

- Single Cloud, Single Region, K8S
- No pet Infrastructure!
- One Cluster per Team can work!
- Cloud automation needs safeguards!

THANK YOU!