When One Line Took Thousands of Websites Offline

Francisco Borges Aurindo Barros, Jack Henschel

Dublin, 2023-10-11

SRECON EUROPE MIDDLE EAST AFRICA

CERN

CERN

- European Organization for Nuclear Research
- Geneva, Switzerland
- Established 1954
- High-energy physics
  – with big machines!

Data Center

- On-premises data center for data acquisition, storage and analysis
- 80% “physics” workload, 20% “online” services

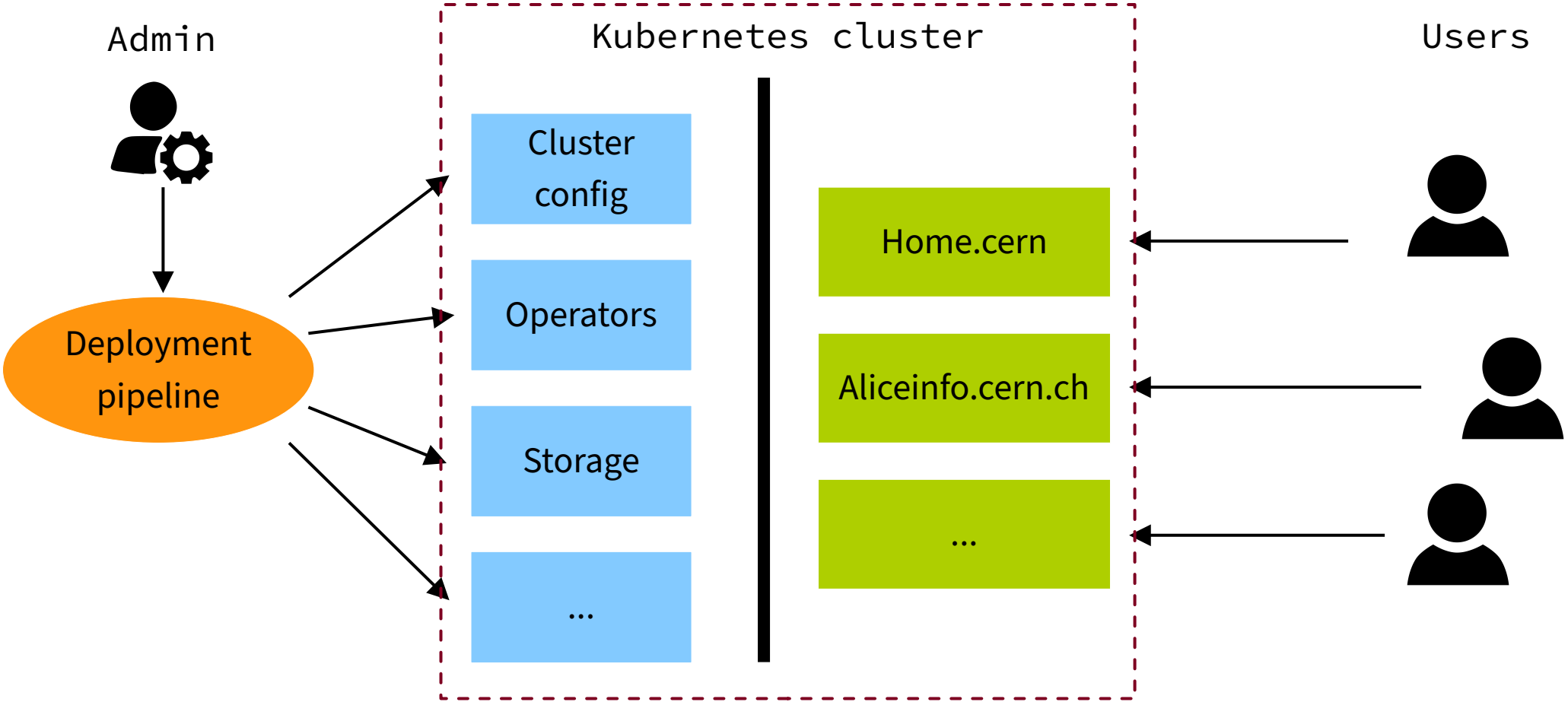
World Wide Web

Web Services Infrastructure on Kubernetes

- Started journey in 2016 with OpenShift Origin 3
- Latest generation built on OKD 4
- Four production clusters
- Today: 8000 web sites/applications/APIs/...
- Small team (5 FTE)
  → lots of automation to keep up with demand

CERN's Journey with OKD: https://youtu.be/6Os9JMNCDXY

Infrastructure and User workloads



Kubernetes Operators

- **Custom Resource Definition (CRD)**:
  - Extends Kubernetes native API
  - OpenAPI schema
- **Custom Resource (CR)**:
  - Concrete object that follows the schema of the CRD
- **Operator**:
  - Custom *controller* that *watches* and *reconciles* the CRs
- → provide a powerful base for self-service SaaS solutions

Operating SaaS at Scale with Operators [KubeConEU'23]: https://youtu.be/0sBgS_3xT8U

```yaml
apiVersion: drupal.webservices.cern.ch/v1alpha1
kind: DrupalSite
metadata:
  name: drupal-tools
spec:
  configuration:
    databaseClass: standard
    diskSize: 1G
    qosClass: standard
    scheduledBackups: enabled
  siteUrl:
    - drupal-tools.web.cern.ch
  version:
    name: v9.4-2
    releaseSpec: RELEASE-2023.02.13T13-47-51Z
status:
  availableBackups: [...]
  dBUpdatesLastCheckTimestamp: 'Feb 14, 2023 at 7:38am (UTC)'
  expectedDeploymentReplicas: 1
```

Managing CERN's fleet of Drupal sites [DrupalCon 2022]: https://youtu.be/uIp_BYTsHq4

CERN Accelerating science

Sign in Directory

ABOUT NEWS SCIENCE RESOURCES SEARCH | EN

... to use unique machines to answer fundamental questions ...

ABOUT NEWS SCIENCE RESOURCES SEARCH | EN

CERN inaugurates Science Gateway, its new outreach centre for science education

CERN has inaugurated its new emblematic centre for science education and outreach targeting a public of all ages. The building was designed by world-renowned architect Renzo Piano and funded by external donations.

7 OCTOBER, 2023

CERN Accelerating science

Sign in Directory

CERN openlab

Training the IT specialists of the future

CERN openlab is a structure designed to create knowledge. We do this through research, development, and evaluation of cutting-edge computing technologies.

This knowledge is disseminated through a wide range of channels, from the publication of reports and articles to the organisation of **lectures and seminars**. By capitalising on our extensive network of connections from research and industry, we are able to secure speakers working at the forefront of new technologies. Wherever possible, we aim to make our lectures and seminars open to all via the Web.

The Incident

Incident Overview

- Initial commit disabling specific Drupal version
- Trigger update of configuration to Kubernetes Cluster

Disable Drupal PHP7 on WebservicesPortal

Changes 1 | Pipelines 3

Showing 1 changed file with 1 addition and 0 deletions

chart/charts/drupal/values.yaml

```
...  ...  @@ -120,4 +120,5 @@ supportedDrupalVersions:
120  120      - "v9.3-1"
121  121      - "v9.3-2"
122  122      - "v8.9-2"
     123  +   - "v9.4-1"
123  124    defaultVersion: "v9.4-2"
```

cluster-integration-tests

- Integration tests and cluster provisioning

configure-cluster

- Configure app-cat-stg cluster
- Configure app-catalogue cluster
- Configure drupal cluster
- Configure drupal-stg cluster
- Configure paas cluster
- Configure paas-stg cluster
- Configure webeos cluster
- Configure webeos-proto cluster
- Configure webeos-stg cluster

- Alerts from monitoring systems and users

Is the `drupal` cluster down?

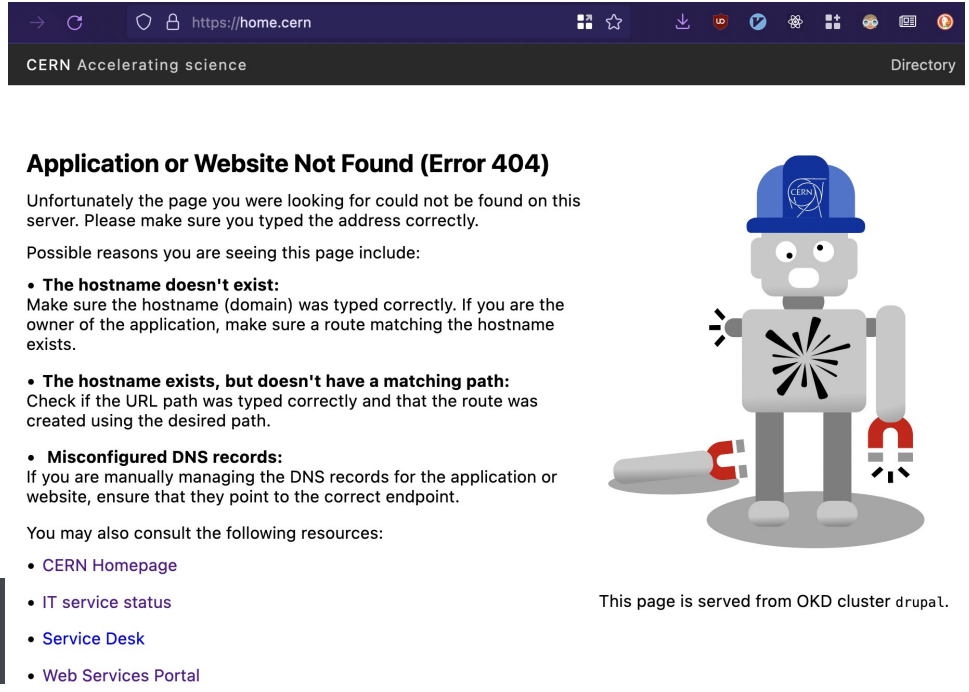
I cannot access any Drupal website right now

https://home.cern/

Down?

Icinga BOT 2:56 PM

🔥 **PROBLEM** home.cern/URLs Drupal Service is WARNING - None



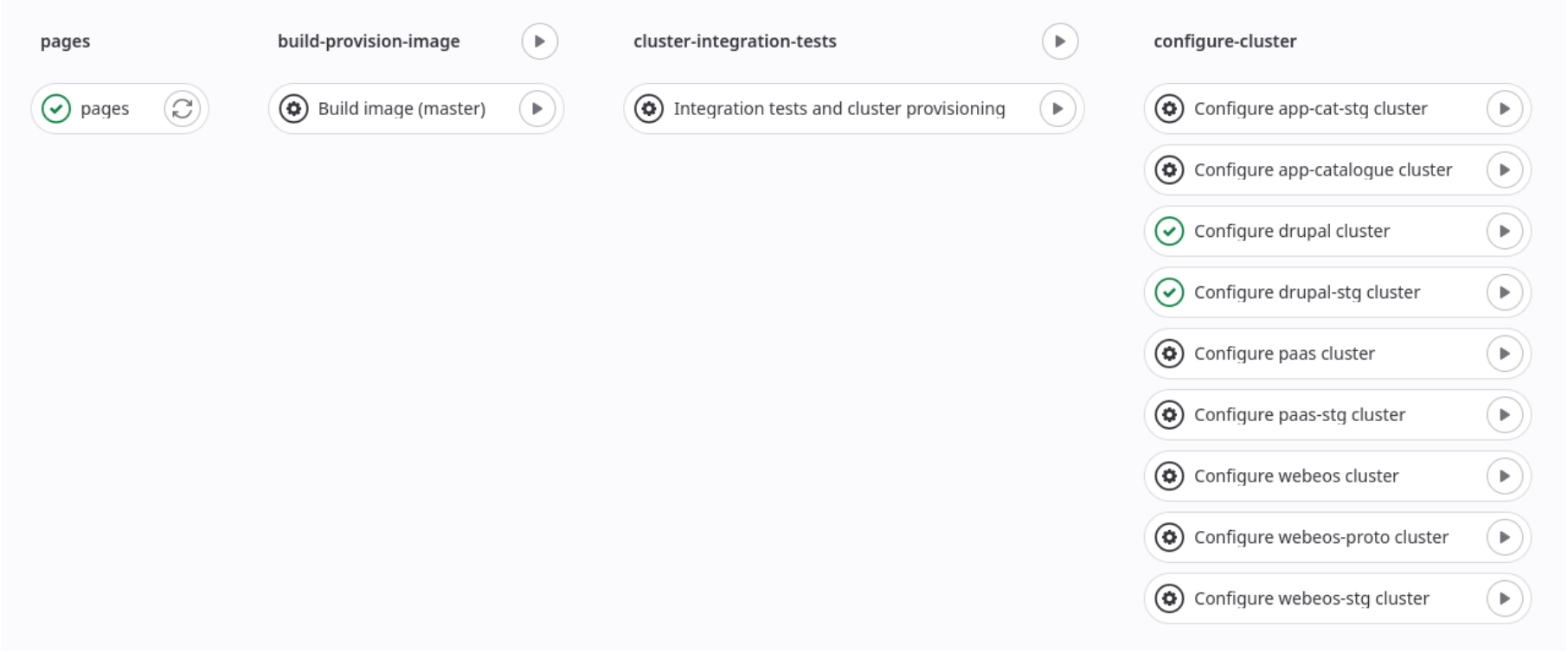
Timeline of events

- 14:44 | **Push to production**
- 14:50 | First **alerts from monitoring and users**
- 15:10 | **Initial assessment** of state
- 15:23 | **Reset cluster configuration** to last working state

(at this point we don't know what happened yet)

Recovering

- **Rollback** infrastructure configuration to previous cluster version



Timeline of events

- 15:25 | **Disable Kubernetes operators**
- 16:40 | Full scale of **outage understood**

Drupal Infrastructure

Internal resources:

Kubernetes manifests

Custom Resources (DrupalSites)

Drupal Cluster

External resources:

CephFS volumes

Authorization API

Databases

Timeline of events

- 15:25 | **Disable Kubernetes operators**
- 16:40 | Full scale of **outage understood**
- 17:00 | **Prioritize recovery procedure** for most important websites
- 21:00 | **home.cern** is back online

(delaying further recovery actions until next day)

Drupal Infrastructure

Internal resources:

Kubernetes manifests

Custom Resources (DrupalSites)

Drupal Cluster

External resources:

CephFS volumes

Authorization API

Database

Recovering

- Restore manifest backups (Velero)

```
velero backup get
velero restore create --from-backup $NAME \
--include-resources-persistentvolumes
```

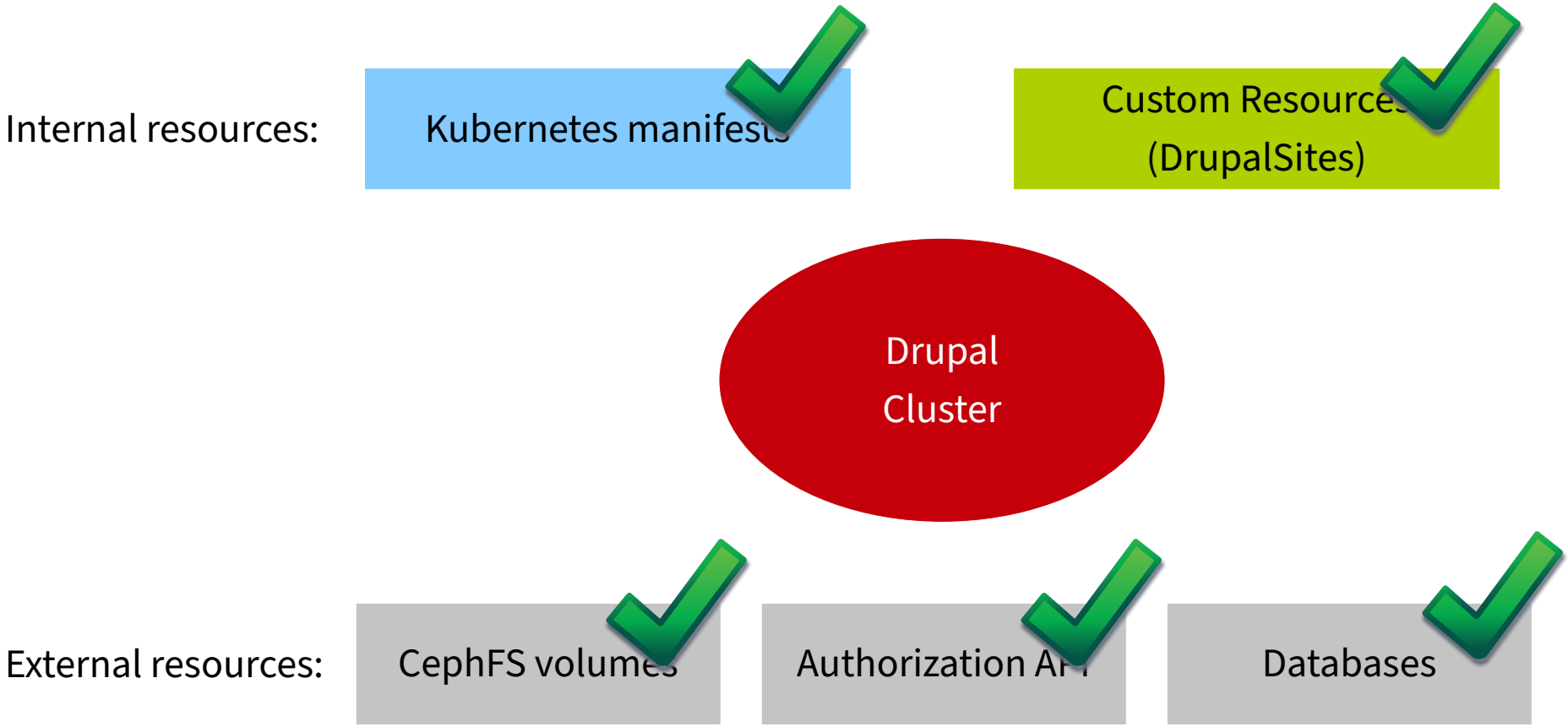
- Re-attach CephFS volumes (soft-deleted with *reclaimPolicy: retain*)

```
kubectl patch pv/$PV_NAME --type json -p '[
  {"op":"remove","path":"/spec/claimRef/uid"},
  {"op":"remove","path":"/spec/claimRef/resourceVersion"},
  {"op":"remove","path":"/metadata/annotations/reclaim-
volumes.cern.ch/volume-reclaim-deletion-timestamp-"}
]'
```

Timeline of events (the next day)

- 9:00 | Prepare and validate procedure to **restore all websites**
- 11:45 | **Request assistance** from DB team for recovery
- 16:00 | All **cluster resources recovered**

Drupal Infrastructure



Root Cause Analysis

- Cluster misconfigured due to a bug in the deployment tool
- Bug introduced on the master branch just before deployment

```
157  167
158  168        files: list[str] = [
159  169            "cluster-defaults.yaml",
160  170            "cluster-id.yaml",
161       -        _secrets_path(cluster_name),
162  171        ]
163  172        if os.path.exists(f"chart/values-{cluster_name}.yaml"):
164  173            files += [f"chart/values-{cluster_name}.yaml"]
165  174        if custom_values_file and os.path.exists(custom_values_file):
166  175            files += [custom_values_file]
167  176
     177  +     files += [_secrets_path(cluster_name)]
     178  +
168  179        for f in files:
```

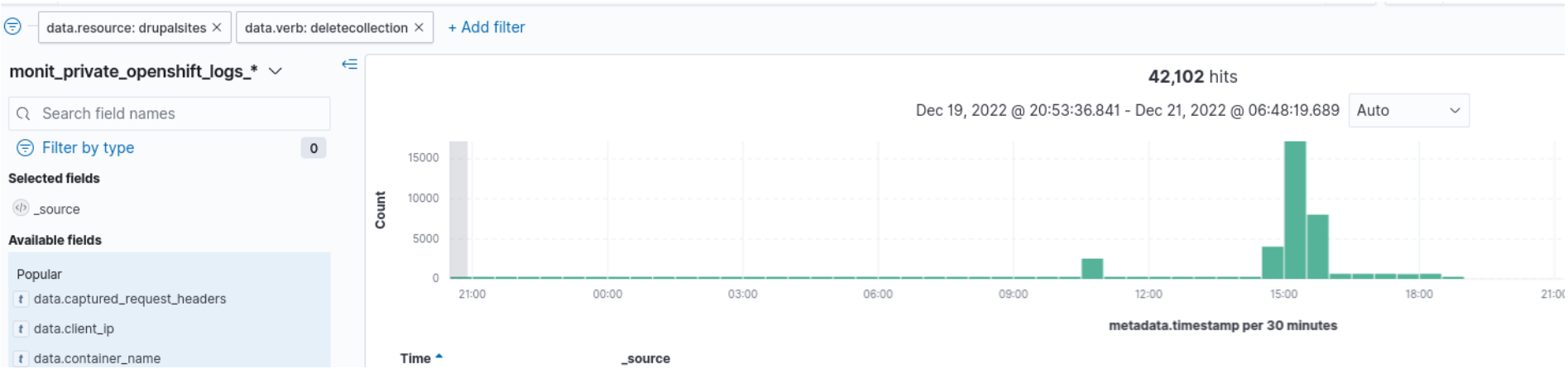
Deployment process

- **End-to-end integration tests** on feature branch
- **Code review** before merging into “master” branch
- Deployment to **staging** environment (triggered by admin)
- **Manual validation** in staging environment
- Deployment to **production** environment
- Internal and external **monitoring** for production

Root Cause Analysis

- “AuthZ” operator drives the **project lifecycle**
- It has the **power to delete projects**
- **Subtle bug** in our deployment tool caused one cluster component to **connect to staging** environment
- Several mitigations in place for invalid state (e.g. no response), but not for *this* case: misconfigured endpoint

Kubernetes Deletion Events



Lessons learned

Mass-deletion and soft-deletion

Deleting** is **easy**, but hard to **undo

- Implement strategies to **delay actual deletion** (*brown-out/scream-test*)
  - Turn off server before decommissioning it
  - Detach volume before deleting data
  - Stop serving website before deleting the content
- Grace period (1 week – 3 months) before final deletion
- If possible: take a backup before deleting
  (How to do that for external resources?)

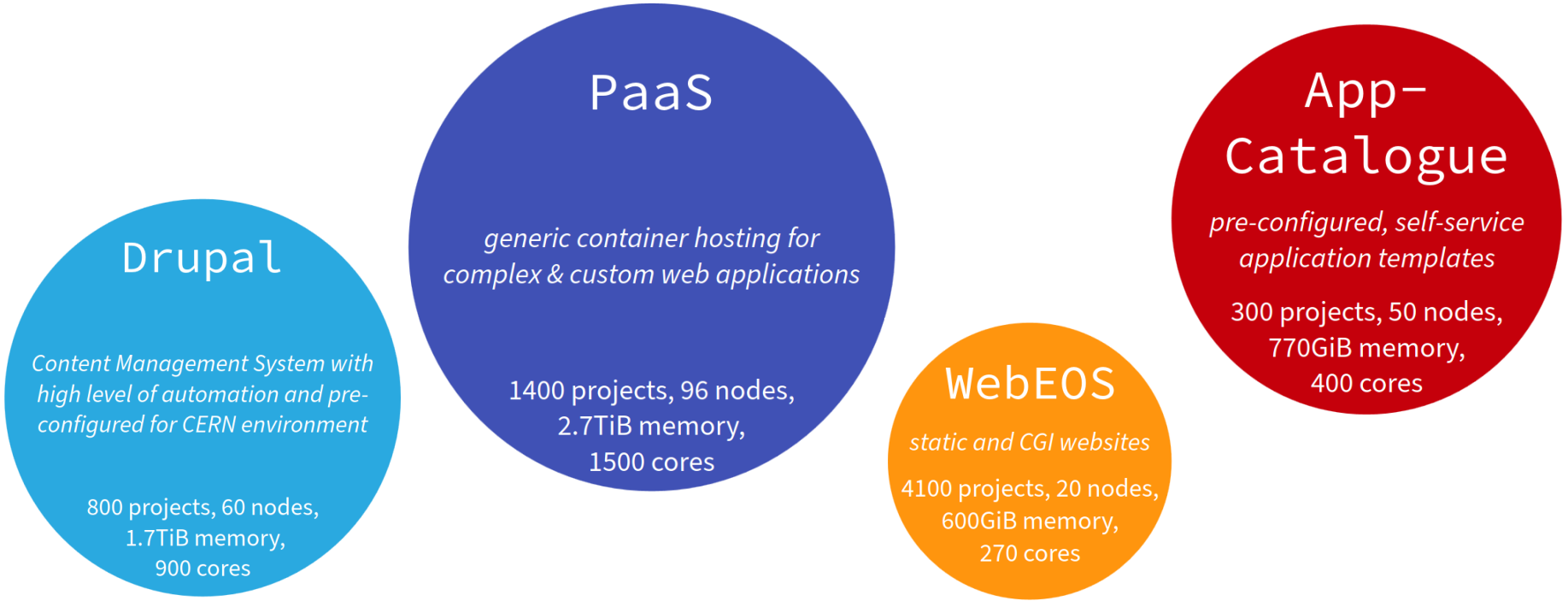
Preview configuration changes

- Code reviews (input) only help so much, also need to verify the output
- Extremely valuable for **confident deployments**
- *tf plan*, *argocd app diff*, *helm diff*, ...

```
@@ -161,7 +162,7 @@
            baselineCapabilitySet: "v4.12"
    path: .
    repoURL: https://gitlab.cern.ch/paas-tools/okd4-deployment/force-clus
-   targetRevision: fb5edd33b1084f1354b151792d489fd468243c10
+   targetRevision: ef85e22f7616d6e5f18493acda865a4807697aea
  syncPolicy:
    automated:
      prune: true
===== argoproj.io/Application openshift-cern-argocd/monitoring-stack-confi
--- /tmp/argocd-diff823853098/monitoring-stack-configuration-live.yaml 202
+++ /tmp/argocd-diff823853098/monitoring-stack-configuration   2023-09-20
@@ -202,7 +202,7 @@
            resources:
              requests:
                cpu: "500m"
-               memory: "3.5Gi"
+               memory: "3500Mi"
            volumeClaimTemplate:
              spec:
                storageClassName: cephfs-no-backup
===== rbac.authorization.k8s.io/ClusterRoleBinding /paas-nviso-okd-audit =
--- /tmp/argocd-diff541664480/paas-okd-audit-live.yaml 2023-09-20 12:35:41
+++ /tmp/argocd-diff541664480/paas-okd-audit   2023-09-20 12:35:41.584964
@@ -1,42 +0,0 @@
-apiVersion: rbac.authorization.k8s.io/v1
-kind: ClusterRoleBinding
-metadata:
-  labels:
```

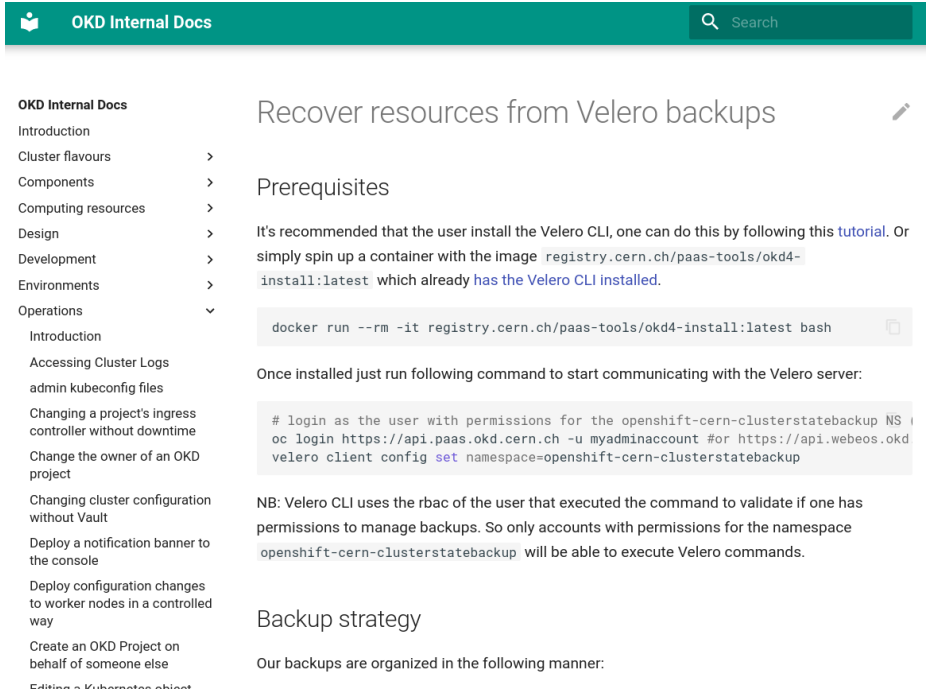
Isolated deployment environments

- Fully isolated clusters prevented the issue from “spreading”



Operational flexibility

- **Reliable**, yet **flexible** disaster recovery procedures
- Requires administrators to be familiar with the tools



Communication channels

- Challenging to handle many sources of input and stakeholders during incident and recovery
- Yet necessary to quickly find mitigations and solutions
- Priorities for recovery should be clear **in advance**
- “War room” participants should be calm and focused on the task

GitOps & Automation

- Fully declarative configuration management for rolling back changes
- Useful to have possibility of pausing automation when needed
- Automations should be able to adopt existing resources

Never underestimate small changes

Thanks to

Abel

Andreas

Carina

David

Joachim

Kate

Lorenzo

Prakhar

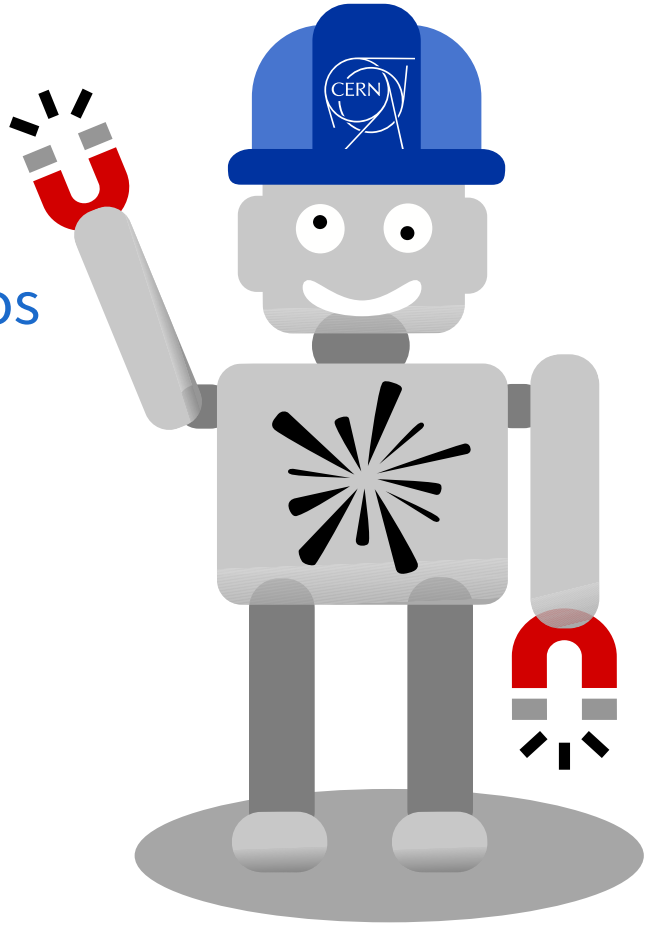
Rajula

When One Line Took Thousands of Websites Offline

Any questions?

Francisco: linkedin.com/in/francisco-aurbarros
Jack: linkedin.com/in/jack-henschel

Slides: https://cern.ch/srecon2023-emea



SRE CON EUROPE MIDDLE EAST AFRICA

Come visit CERN!