When One Line Took Thousands of Websites Offline

Francisco Borges Aurindo Barros, Jack Henschel

Dublin, 2023-10-11
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

Deployment pipeline

Admin

Kubernetes cluster

Cluster config
Operators
Storage
...

Home.cern
Aliceinfo.cern.ch
...

Users
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*
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 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
The Incident
Incident Overview

- Initial commit disabling specific Drupal version
- Trigger update of configuration to Kubernetes Cluster
• Alerts from monitoring systems and users

Is the drupal cluster down?

I cannot access any Drupal website right now

https://home.cern/ Down?

Application or Website Not Found (Error 404)

Unfortunately the page you were looking for could not be found on this server. Please make sure you typed the address correctly.

Possible reasons you are seeing this page include:

• The hostname doesn’t exist:
  Make sure the hostname (domain) was typed correctly. If you are the owner of the application, make sure a route matching the hostname exists.

• The hostname exists, but doesn’t have a matching path:
  Check if the URL path was typed correctly and that the route was created using the desired path.

• Misconfigured DNS records:
  If you are manually managing the DNS records for the application or website, ensure that they point to the correct endpoint.

You may also consult the following resources:

• CERN Homepage
• IT service status
• Service Desk
• Web Services Portal

This page is served from OKD cluster drupal.
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)**
  
  ```bash
  velero backup get
  velero restore create --from-backup $NAME \
  --include-resources-persistentvolumes
  ```

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

Internal resources:
- Kubernetes manifests
- Custom Resources (DrupalSites)

Drupal Cluster

External resources:
- CephFS volumes
- Authorization API
- Databases
Root Cause Analysis

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

```python
files: list[str] = [
    "cluster-defaults.yaml",
    "cluster-id.yaml",
    _secrets_path(cluster_name),
]

if os.path.exists(f"chart/values-{cluster_name}.yaml"):
    files += [f"chart/values-{cluster_name}.yaml"]

if custom_values_file and os.path.exists(custom_values_file):
    files += [custom_values_file]

files += [_secrets_path(cluster_name)]

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

Deletions 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`, ...

```yaml
baselineCapabilitySet: "v4.12"
path: .
reposURL: https://gitlab.cern.ch/paas-tools/okd4-deployment/force-clustering
- targetRevision: b93edd338b19e4f135b151792d489ff4d6243c10
  targetRevision: ef85e22f7466e9f18493a9d865a18076797ae9
syncPolicy:
  automated: true
  prune: true

--- argoproj.io/Application openshift-cern-argocd/monitoring-stack-conf
--- /tmp/argocd-diff823853098/monitoring-stack-configuration-live.yaml 2023-09-20
+++ /tmp/argocd-diff823853098/monitoring-stack-configuration-live.yaml 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-diff541664486/paas-okd-audit-cluster-binding-live.yaml 2023-09-20 12:35:4
+++ /tmp/argocd-diff541664486/paas-okd-audit-cluster-binding-live.yaml 2023-09-20 12:35:41.59464
@@ -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”

**Drupal**
- Content Management System with high level of automation and pre-configured for CERN environment
- 800 projects, 60 nodes, 1.7TIB memory, 900 cores

**PaaS**
- generic container hosting for complex & custom web applications
- 1400 projects, 96 nodes, 2.7TIB memory, 1500 cores

**App-Catalogue**
- pre-configured, self-service application templates
- 300 projects, 50 nodes, 770GiB memory, 400 cores

**WebEOS**
- static and CGI websites
- 4100 projects, 20 nodes, 600GiB memory, 270 cores
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
Any questions?

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

Slides: https://cern.ch/srecon2023-emea
Come visit CERN!