

Effectively manage production incidents

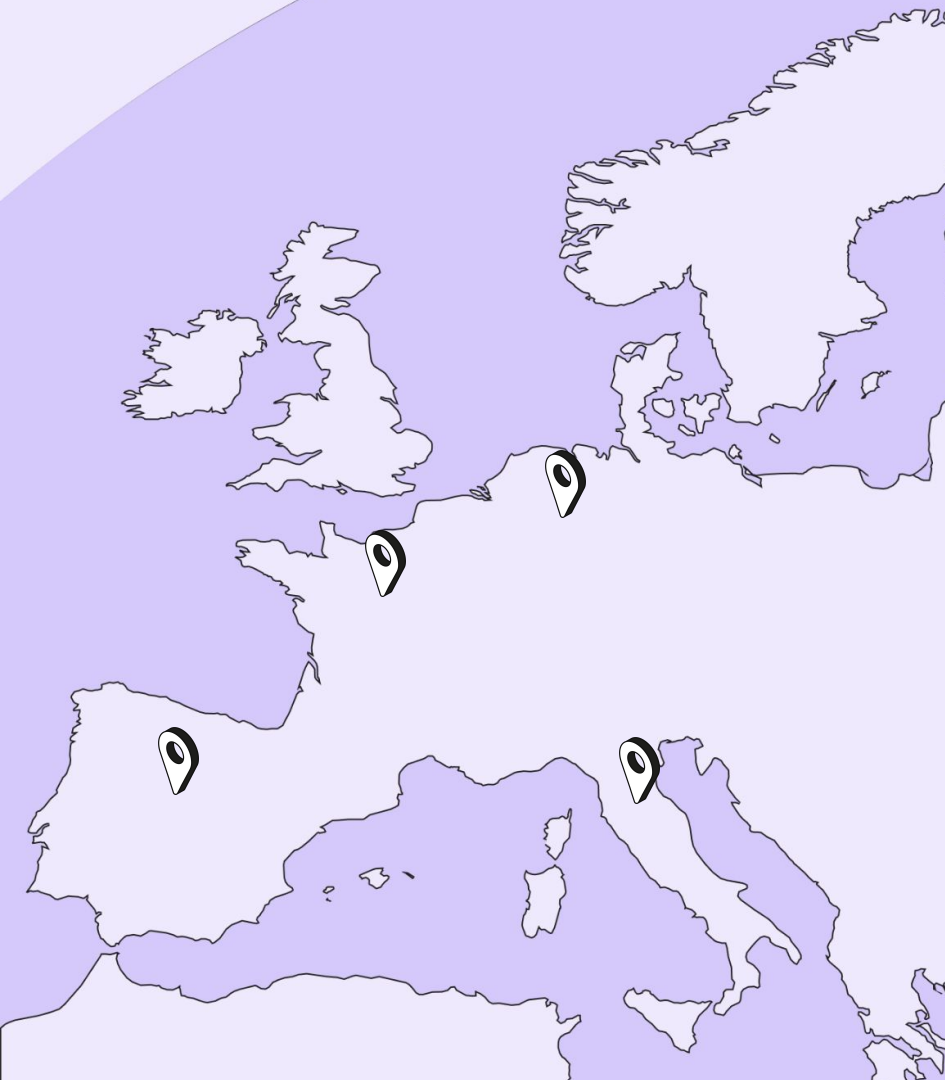
Mathieu Corbin

Senior staff SRE at Qonto

<https://www.mcorbin.fr/>

@_mcorbin





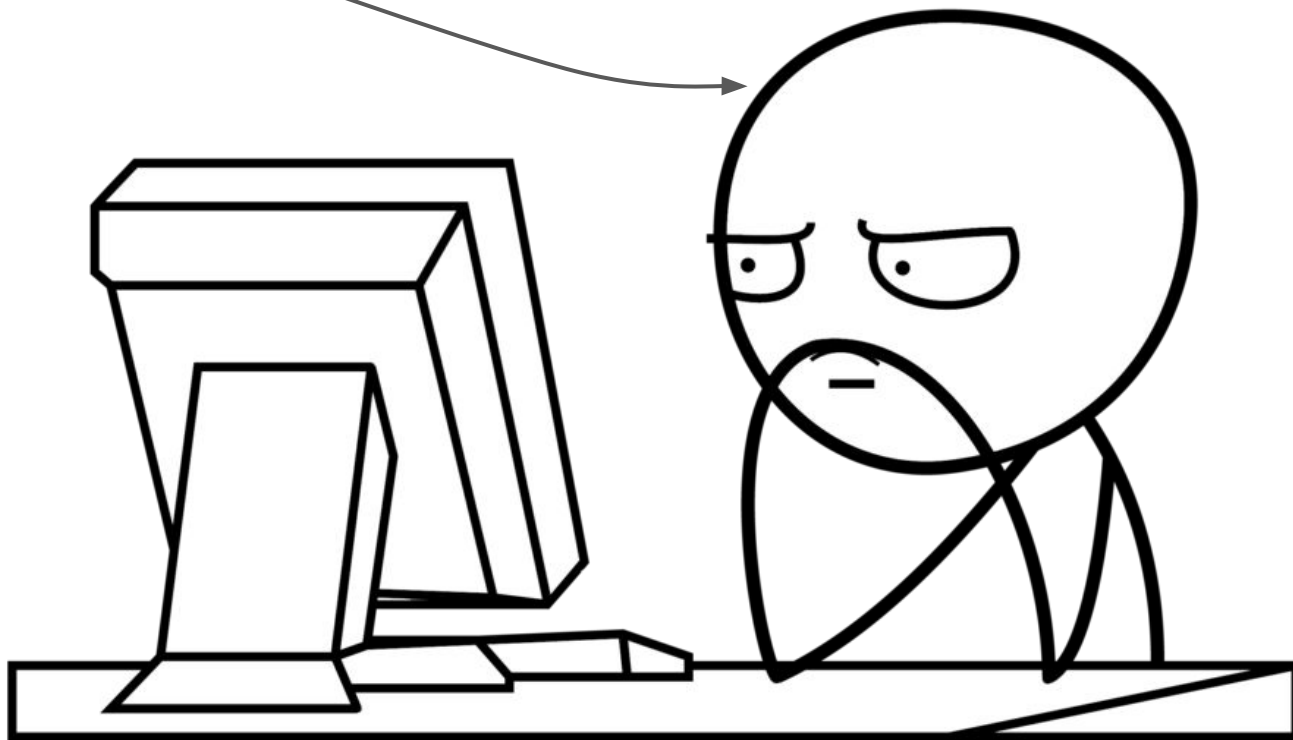
Qonto

All your business finances
managed in one app



<https://qonto.com/>

SRE (me) at work



- #97564 **P3** [KubePodNotReady] Pod overprovisioner/overprovisioner-744bb945d6-4h72h has been in a non-ready state for longer than an hour.
x2
staging
👤 SRE

- #97565 **P3** [KubePodNotReady] Pod overprovisioner/overprovisioner-744bb945d6-mrwn5 has been in a non-ready state for longer than an hour.
x2
staging
👤 SRE

- #97563 **P3** [ArgoCDApplicationError] 40 ArgoCD <https://gitlab.qonto.co/qonto/qonto-backoffice> could not sync for more than 20m
x7
staging
👤 Onboarding

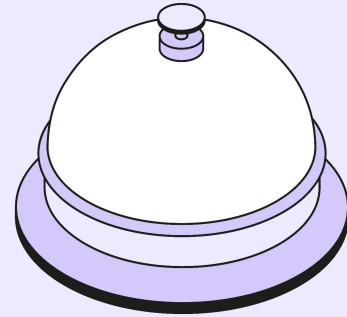
- #97562 **P3** [KubeDeploymentReplicasMismatch] Deployment kube-system/external-secrets-kubernetes-external-secrets has not matched the exp...
x2
staging
👤 SRE

Symptoms

- Applications cannot be deployed anymore on a staging cluster
 - Lot of pods (including “overprovisioner” pods) cannot be scheduled
 - Pods crashing (external secret operator down)
 - Requests throttling in kubernetes-related components (ArgoCD)


Declaring an incident

Everyone can do it at any time from Slack



Declaring an incident

#incident-report slack channel
/incident

 **Incident form** 📄 ✕

Incident name

Incident description

 **Incident automation** APP 12:59 PM
New incident reported by @Mathieu Corbin

Staging: Kubernetes cluster 3 is misbehaving: We have performance issues and pods crashing on the staging 3 kubernetes cluster

Slack channel

#2023-03-27-staging-kubernetes-cluster-3-is-misbehaving



Report

Incident report

Declaring an incident

2023-03-27-staging-kubernetes-cluster-3-is-misbehaving  <https://www.notion.so/Staging-Kubernetes-cluster-3-is-misbehaving-70d037b488c04f5e819e62c669dac715>

 1 Pinned

Tip: Try  to search this channel 

2023-03-27-staging-kubernetes-cluster-3-is-misbehaving

 @Incident automation created this channel on March 27th. This is the very beginning of the # 2023-03-27-staging-kubernetes-cluster-3-is-misbehaving channel.

Monday, March 27th 



Incident automation APP 12:59 PM

joined #2023-03-27-staging-kubernetes-cluster-3-is-misbehaving. Also, Mathieu Corbin joined.



Incident automation APP 12:59 PM

set the channel topic: <https://www.notion.so/Staging-Kubernetes-cluster-3-is-misbehaving-70d037b488c04f5e819e62c669dac715>



Incident automation APP 12:59 PM

Incident commander: [@Mathieu Corbin](#)

Use this Slack channel to assemble the response team and communicate the incident's resolution.

Dedicated incident commander channel: [#2023-03-27-incident-commander-staging-kubernetes-cluster-3-is-misbehaving](#)

 Slack huddle or use [meet.new](#).

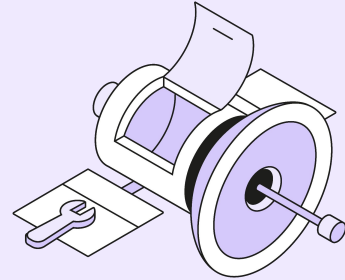
Did something happen recently? [production](#) / [staging](#)

Declaring an incident

- One **slack channel** and one **Notion page** automatically created for the incident
 - People then join the slack channel/meet
- One incident commander designed
 - Coordination (priority: **mitigate**)
 - Handle communication
 - Make sure everything is clear for everyone
 - Ask questions, ask to reformulate if needed
 - Drive and prioritize actions
 - Complete the Notion page and is responsible for the follow-up

Notion page

Centralize information about the incident





Staging: Kubernetes cluster 3 is misbehaving

📄 *Status	4 - post-mortem completed
📅 *Start/end dateti...	March 27, 2023 12:59 → 14:13
Σ Time to resolve	1hr 14min
☰ *Estimated numb...	0
Σ *Customer Impact	XL - Over 15% of customers
📄 *Single-customer...	S - no or slight disturbance for the customer (see examples below)
☰ *Money at risk (€)	0
# *Number of Trans...	0
Σ Severity	🟡 SEV-3
👤 *Incident comma...	👤 Mathieu Corbin
👤 *Response team	
📄 *Detected via	Automatic alert (OpsGenie or other tools)
🔗 *Concerned Thir...	Empty
🔗 *Root Cause	📄 Internal - System failure - Failure of IT systems in production
☰ *Features impact...	None
☰ *Countries impac...	Does not apply
☰ Department	Empty
☰ CFT	Empty
☰ Stack	SRE Backend / Platform
☰ Tech team	SRE Platform Reliability
🕒 Last edited time	September 1, 2023 1:49 PM
🔗 *Slack channel U...	https://qonto.slack.com/archives/C050DFJ932P
🔗 🌐 [DB] PDCA	📄 Staging: Kubernetes cluster 3 is misbehaving

▶ 🖱️ How to fill this page - for the incident commander

🗨️ Communications & Emergency Contacts

We are on Meet

▶ Need to contact Provider

▶ Impact on cards service?

We are on Slack Huddle

▶ Need Tech on-duty

▶ R&C escalation

▶ Comms to clients

📄 Summary

Description - what is happening?

The staging3 kubernetes cluster is misbehaving (timeouts, pods crashing...): a lot of features branches on this cluster are not working anymore.

Cause - why did it happen?

- Crossplane was installed manually on the cluster and create latency on the kubernetes API server + made external-secrets crash
- We reached the maximum number of nodes in the staging cluster3.

Impact - how bad is it?

Several developers had issues with their feature branches environments|

Mitigation - how did we react?

Deleting crossplane + scale the cluster

Timeline

Europe/Paris timezone

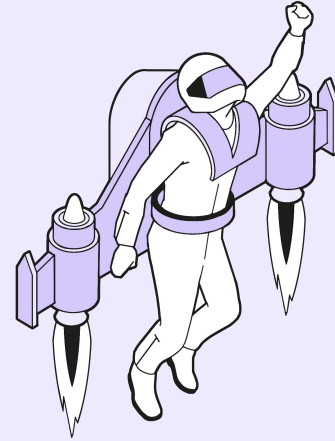
2023-03-27

- **12:59** : incident created due to several alerts on this cluster + duty complains from developers.
 - Pods not being provisioned
 - External secret controller crashing
 - The cluster 3 is immediately disabled from branches placement API
- **13:10** : after investigations we found that crossplane was installed on this cluster and seems to be the root cause of the slowness. We decide to delete it (delete crossplane namespace + delete the ~1000 crossplane CRD installed)
- **13:44** : everything is deleted except 3 CRDs that are stuck
- **14:00** : we noticed that the AWS ASG for cluster3 are full (60/60 instances deployed): we scale them to 80
- **14:10** : the situation is now stable

+ All the information/data that are relevant (link to logs, graphs, screenshots...)

PDCA

Once the incident is mitigated



PDCA: **P**lan **D**o **C**heck **A**ct


- Frame the problem (problem statement)
- Describe what happened (incident timeline usually)
- Find incidents root causes
- **Do** actions to resolve them
 - Should be **simple actions** (timeboxed)
 - **Check** them
 - **Act**: Apply them to the rest of the systems/company

Aa Name	🕒 Created time	👤 Owner	👥 Participants	☀️ Status	📄 CFT	🔗 Incident
 Staging: Kubernetes cluster 3 is misbehaving	2023/04/06 3:01 PM	 Mathieu Corbin		● Plan done		 Staging: Kube cluster 3 is misbe


+ New

New notion page

Problem statement

 **Problem statement:** The staging3 kubernetes cluster is misbehaving (timeouts, pods crashing and not scheduling...): a lot of features branches on this cluster are not working anymore.

Context

 If it comes from an incident, copy-paste the timeline.
If it comes from a red bin or a QA return, copy-paste the context.

Root cause analysis

The most difficult part of the PDCA ?

[Occurrence] Why did the problem occur?

- The Kubernetes control plane API server was lagging and this caused some kubernetes controllers to crash: (for example: without internal secrets, applications cannot be deployed anymore). Why ?
 - Crossplane was installed on the staging3 for testing purposes and its installation (890 CRDs) caused the API server to generate timeouts and Prometheus to OOMkill (cardinality issue). Why ?
 - [RC 2] Crossplane was deployed directly on staging to test its integration with a feature branch environment.

[Non-detection] Why didn't we detect it sooner?

- We hit the maximum size of the general-purpose AWS autoscaling group: this prevented new pods to be scheduled. Why ?
 - We had a lot branches deployed on this cluster and we didn't detected that we hit the limit before it was too late. Why ?
 - [RC 1] We introduced a regression a few months ago on the metric that prevent us to reach 100%
 - PR to fix it

[DB] Do & Check

☰ Root cause

Aa Countermeasure

👤 Owner

▼ Depart...

▼ Tech st...

▼ CFT

▼ Tech team

[RC 1]

📄 Fix the alert for the problematic AWS autoscaling group

👤 Mathieu Corbin

SRE Platform Reliability

[RC 2]

📄 Allow only some CRDs to be deployed

👤 Mathieu Corbin

SRE Platform Reliability

+ New

COUNT 2

☰ Check

📅 Check date

⚙️ Status

↗️ PDCA

↗️ 📄 [DB] Act

🕒 Created time

The alert was executed in staging to see its result

March 28, 2023

● Check OK

📄 Staging: Kubernetes cluster 3 is misbehaving

📄 Make sure that this alert exists and is correctly defined in all environments/all ASG

April 6, 2023 3:19 PM

Create a CRD not in the list ⇒ it should be blocked

March 28, 2023

● Check OK

📄 Staging: Kubernetes cluster 3 is misbehaving

📄 Deploy the configuration on all environments

April 6, 2023 3:27 PM

Conclusion

- Clear procedures for incidents management
- Easy to use tooling
 - Slack
 - Notion templates
- Continuous improvements
 - Problem => PDCA (repeat)
 - “Deep” issues => Kaizen, A3
- Doing good PDCA (right root causes, counter measures) **requires training !**

Thank you !



Our blog

<https://medium.com/qonto-way/tagged/tech>



Jobs

<https://qonto.com/en/careers>



Questions ?

contact@mcobin.fr