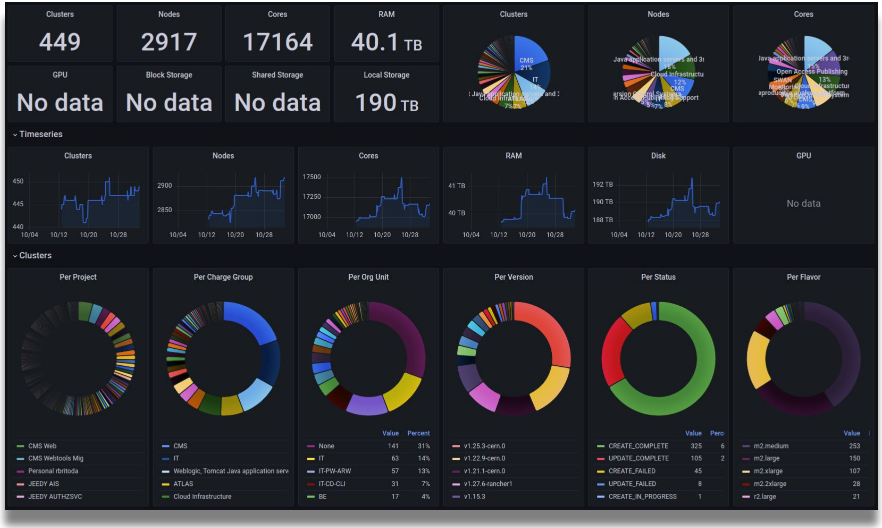
Kubernetes @ CERN

Spyros Trigazis, IT-PW

Meeting with Université de Lausanne

https://indico.cern.ch/event/1417842/

Kubernetes Service



Clusters 449 | Nodes 2917 | Cores 17164 | RAM 40.1 TB

GPU No data | Block Storage No data | Shared Storage No data | Local Storage 190 TB

Per Project
CMS Web
CMS Webtools Mig
Personal rbritoda
JEEDY AIS
JEEDY AUTHZSVC

Per Charge Group
CMS
IT
Weblogic, Tomcat Java application serv
ATLAS
Cloud Infrastructure

	Value	Percent
None	141	31%
IT	63	14%
IT-PW-ARW	57	13%
IT-CD-CLI	31	7%
BE	17	4%

Per Version
v1.25.3-cern.0
v1.22.9-cern.0
v1.21.1-cern.0
v1.27.6-rancher1
v1.15.3

	Value	Perc
CREATE_COMPLETE	325	6
UPDATE_COMPLETE	105	2
CREATE_FAILED	45	
UPDATE_FAILED	8	
CREATE_IN_PROGRESS	1	

	Value
m2.medium	253
m2.large	150
m2.xlarge	107
m2.2xlarge	28
r2.large	21

On Demand Cluster as a Service

Multiple versions, custom features

Integration with CERN networking,

identity, security, storage, …

Used by multiple applications with resources in 513 across most sectors:

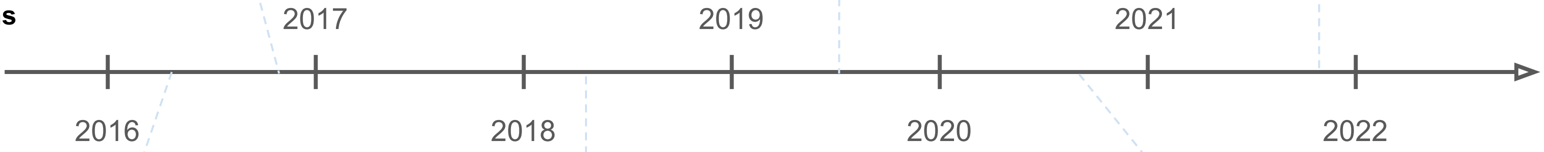
IT (Registry, GitLab + CI, MONIT, SWAN, Kubeflow/ML, SSO, CS/IP1, …)

IT / FHR (EDH, EDMS, Phonebook, AIS, egroups, Learning Hub, …)

RCS (ATLAS Rucio, CMSWeb, InspireHEP, HEPData, …)

Clusters	20	122	276	342	330	334	411
Nodes		700	1098	1093	1817	2560	2780

Investigations



Pilot Service

Kubernetes, Swarm

No obvious around this time which orchestrator would win. **Scalability and performance** tests.

“Production” Service

Kubernetes, Swarm, **Mesos**

Integration with the CERN and WLCG environments. Certificates, Storage.

CephFS, Manila, CVMFS

CSI everywhere. Developed initial drivers for **CephFS and Manila**. Improvements for CVMFS.

Work done **fully upstream**, taken later by multiple other companies.

High Availability, GitOps, Public Cloud

GitOps and **Secret Management**, Dissemination. Helm and Flux/ArgoCD.

HA improvements with **Node Groups**, **Cluster Auto Scaling**, **Auto Healing**, **LBaaS** for serviceType: LB.

Dissemination, Security

Container Webinars on infrastructure and use cases, popular elsewhere as well.

Re-thinking with Security Team **security aspects of containerized deployments**, policies, best practices.

Disaster Recovery

Volume Snapshots, Automated Backups and Restore.

Round one and multiple items to work together with the community.

Integration: Storage

Physics Data - EOS, initially eosxd with hostPath mount, then eosxd-csi

General Purpose backed by CEPH - CSI plugins: manila, cinder, cephfs

Software distribution - csi-cvmfs POSIX read-only via FUSE

Other, TN - csi-driver-nfs for NetAPP or custom NFS servers

Integration: Networking & Load Balancing

Networking

- Calico is the default CNI
- Cilium is opt-in, attractive for cluster-mesh, hybrid deployments

Load Balancing

- Ingresses with DNS alias
- LoadBalancer via OpenStack/Octavia

Integration: Certificates

Let’s Encrypt, self-service with cert-manager.io

- HTTP-01 challenge, LE allowed in perimeter firewall (upon request)
- DNS-01 challenge, cert-manager and CERN DNS integration

CERN CA, custom daemonset

- Host Certificate per node

Integration: Monitoring

Metrics

- Upstream kube-prometheus-stack
- WIP aggregation to central infrastructure

Logging

- Fluentd daemonset with http plugin, pushing to central gateway
- Transition to fluentbit

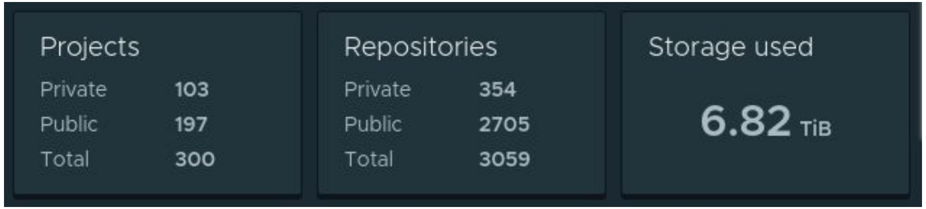
Operations: Registry

Based on goharbor

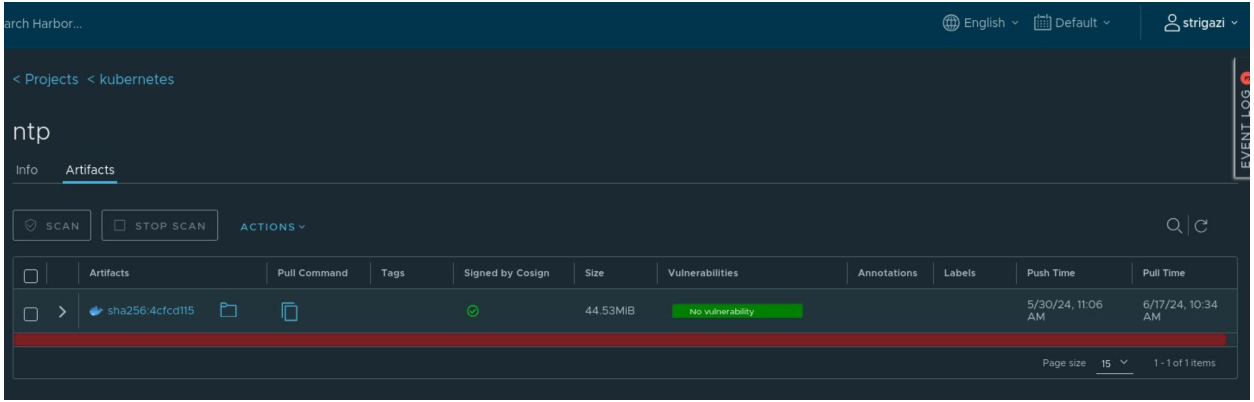
Integration with gitlab-ci

Data in s3 by ceph

Vulnerability scanning via trivy.dev



Projects		Repositories		Storage used
Private	103	Private	354	6.82 TiB
Public	197	Public	2705	
Total	300	Total	3059	



< Projects < kubernetes

ntp

□		Artifacts	Pull Command	Tags	Signed by Cosign	Size	Vulnerabilities	Annotations	Labels	Push Time	Pull Time
□	>	sha256:4cfcd115			✓	44.53MiB	No vulnerability			5/30/24, 11:06 AM	6/17/24, 10:34 AM

Page size 15 1 - 1 of 1 items

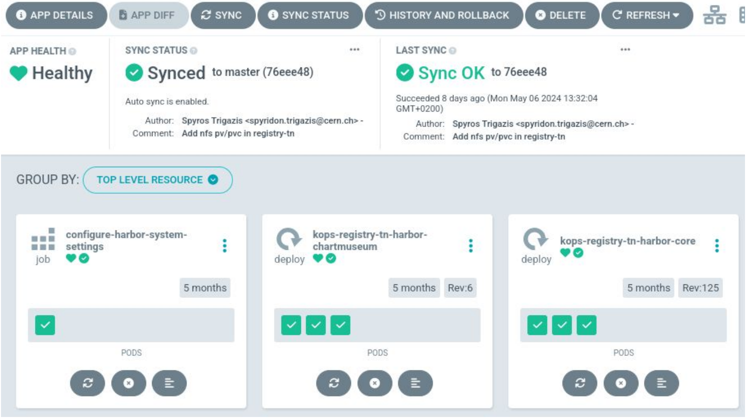
Operations: ArgoCD

Central repo (private) for all applications

Each application set manage in public repo eg registry

Integration with private vault for secrets

Alerts via grafana to mail mattermost, telegram



APP HEALTH
Healthy

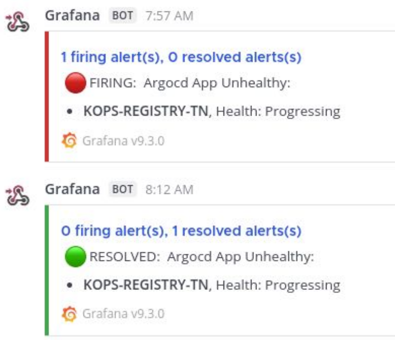
SYNC STATUS
Synced to master (76eee48)
Auto sync is enabled.
Author: Spyros Trigazis <spyridon.trigazis@cern.ch> -
Comment: Add nfs pv/pvc in registry-tn

LAST SYNC
Sync OK to 76eee48
Succeeded 8 days ago (Mon May 06 2024 13:32:04 GMT+0200)
Author: Spyros Trigazis <spyridon.trigazis@cern.ch> -
Comment: Add nfs pv/pvc in registry-tn

configure-harbor-system-settings
job
5 months

kops-registry-tn-harbor-chartmuseum
deploy
5 months Rev:6

kops-registry-tn-harbor-core
deploy
5 months Rev:125



Grafana BOT 7:57 AM

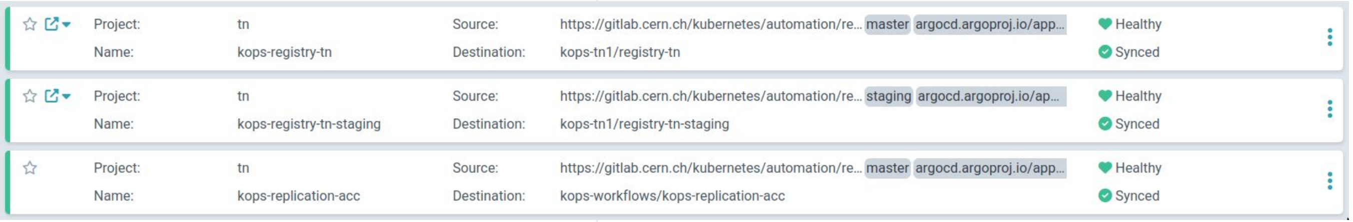
1 firing alert(s), 0 resolved alert(s)
FIRING: Argocd App Unhealthy:
- KOPS-REGISTRY-TN, Health: Progressing

Grafana v9.3.0

Grafana BOT 8:12 AM

0 firing alert(s), 1 resolved alert(s)
RESOLVED: Argocd App Unhealthy:
- KOPS-REGISTRY-TN, Health: Progressing

Grafana v9.3.0



Project:	tn	Source:	https://gitlab.cern.ch/kubernetes/automation/re... master argocd.argoproj.io/app...	Healthy
Name:	kops-registry-tn	Destination:	kops-tn1/registry-tn	Synced
Project:	tn	Source:	https://gitlab.cern.ch/kubernetes/automation/re... staging argocd.argoproj.io/ap...	Healthy
Name:	kops-registry-tn-staging	Destination:	kops-tn1/registry-tn-staging	Synced
Project:	tn	Source:	https://gitlab.cern.ch/kubernetes/automation/re... master argocd.argoproj.io/app...	Healthy
Name:	kops-replication-acc	Destination:	kops-workflows/kops-replication-acc	Synced

Work in progress

BC/DR

Multi-cluster / cluster mesh

Improved audit logging with Falco

Adoption of CAPI for cluster lifecycle

SBOM integration for containers

Links

https://gitlab.cern.ch/kubernetes/magnum

https://gitlab.cern.ch/kubernetes/automation/releases/cern-magnum

https://gitlab.cern.ch/kubernetes/networking/

https://gitlab.cern.ch/kubernetes/security/

Q & A