Dynamic storage provisioning for elastic cloud services with dCache

International Symposium on Grids & Clouds (ISGC) 2021

Michael Schuh, Patrick Fuhrmann, Tigran Mkrtchyan, Paul Millar, Johannes Reppin, Christian Voss, Tim Wetzel

March 26 2021

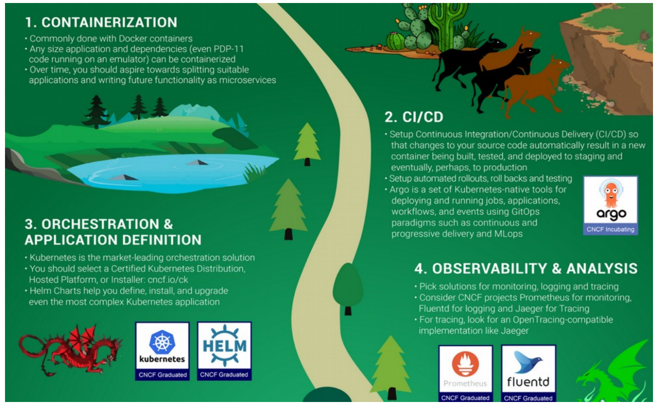
The Cloud Native Landscape

The Cloud Native Trail Map

Cloud environments for scientific use-cases

- public, private, hybrid

1. Containers, Microservices
2. Robust automation
3. Orchestration
4. Monitoring / Analysis



1. CONTAINERIZATION

- Commonly done with Docker containers
- Any size application and dependencies (even PDP-11 code running on an emulator) can be containerized
- Over time, you should aspire towards splitting suitable applications and writing future functionality as microservices

2. CI/CD

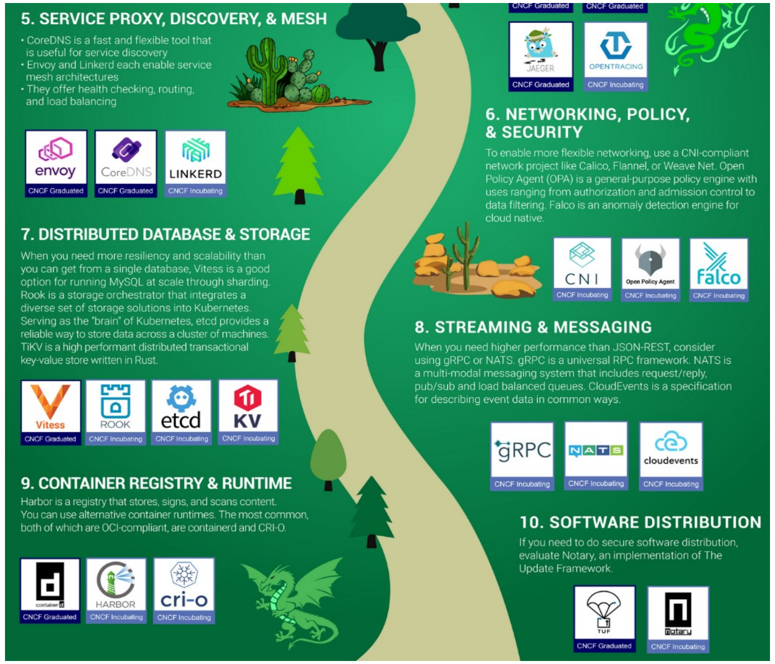
- Setup Continuous Integration/Continuous Delivery (CI/CD) so that changes to your source code automatically result in a new container being built, tested, and deployed to staging and eventually, perhaps, to production
- Setup automated rollouts, roll backs and testing
- Argo is a set of Kubernetes-native tools for deploying and running jobs, applications, workflows, and events using GitOps paradigms such as continuous and progressive delivery and MLops

3. ORCHESTRATION & APPLICATION DEFINITION

- Kubernetes is the market-leading orchestration solution
- You should select a Certified Kubernetes Distribution, Hosted Platform, or Installer: cncf.io/ck
- Helm Charts help you define, install, and upgrade even the most complex Kubernetes application

4. OBSERVABILITY & ANALYSIS

- Pick solutions for monitoring, logging and tracing
- Consider CNCF projects Prometheus for monitoring, Fluentd for logging and Jaeger for Tracing
- For tracing, look for an OpenTracing-compatible implementation like Jaeger



5. SERVICE PROXY, DISCOVERY, & MESH

- CoreDNS is a fast and flexible tool that is useful for service discovery
- Envoy and Linkerd each enable service mesh architectures
- They offer health checking, routing, and load balancing

6. NETWORKING, POLICY, & SECURITY

To enable more flexible networking, use a CNI-compliant network project like Calico, Flannel, or Weave Net. Open Policy Agent (OPA) is a general-purpose policy engine with uses ranging from authorization and admission control to data filtering. Falco is an anomaly detection engine for cloud native.

7. DISTRIBUTED DATABASE & STORAGE

When you need more resiliency and scalability than you can get from a single database, Vitess is a good option for running MySQL at scale through sharding. Rook is a storage orchestrator that integrates a diverse set of storage solutions into Kubernetes. Serving as the "brain" of Kubernetes, etcd provides a reliable way to store data across a cluster of machines. TiKV is a high performant distributed transactional key-value store written in Rust.

8. STREAMING & MESSAGING

When you need higher performance than JSON-REST, consider using gRPC or NATS. gRPC is a universal RPC framework. NATS is a multi-modal messaging system that includes request/reply, pub/sub and load balanced queues. CloudEvents is a specification for describing event data in common ways.

9. CONTAINER REGISTRY & RUNTIME

Harbor is a registry that stores, signs, and scans content. You can use alternative container runtimes. The most common, both of which are OCI-compliant, are containerd and CRI-O.

10. SOFTWARE DISTRIBUTION

If you need to do secure software distribution, evaluate Notary, an implementation of The Update Framework.

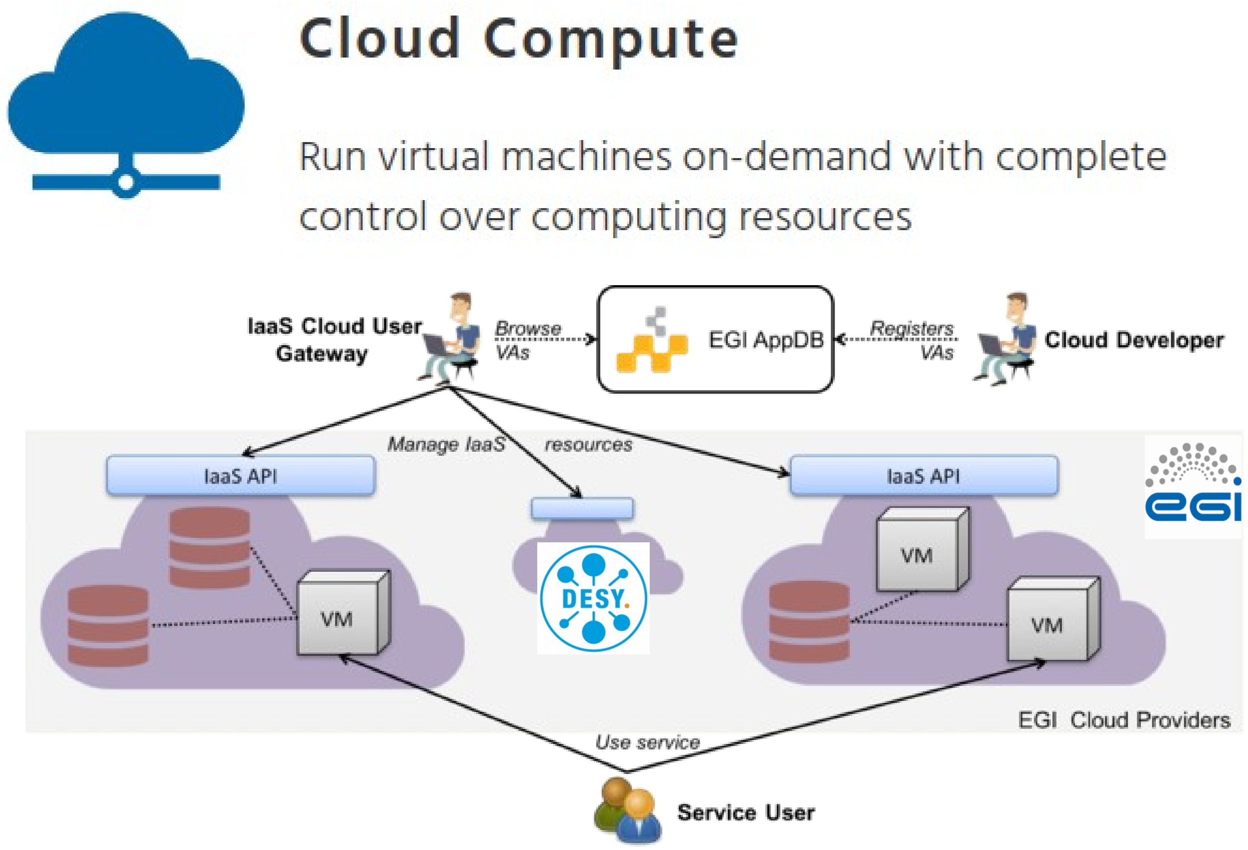
Integrate with research infrastructure

5. DNS
6. Network Operations
   - LBaaS
   - Dynamic Certificates
7. Scientific Data, storage
   - **dCache**
   - High performance storage
8. Event streaming platforms
   - Data Acquisition streams
   - FaaS
9. Scale container registry
   - HPC / HTC
10. Software Repository
    - CVMFS

Source: trailmap.cncf.io

The EGI Federated Cloud

DESY provides resources to the EGI Federated Cloud



Cloud Compute

Run virtual machines on-demand with complete control over computing resources

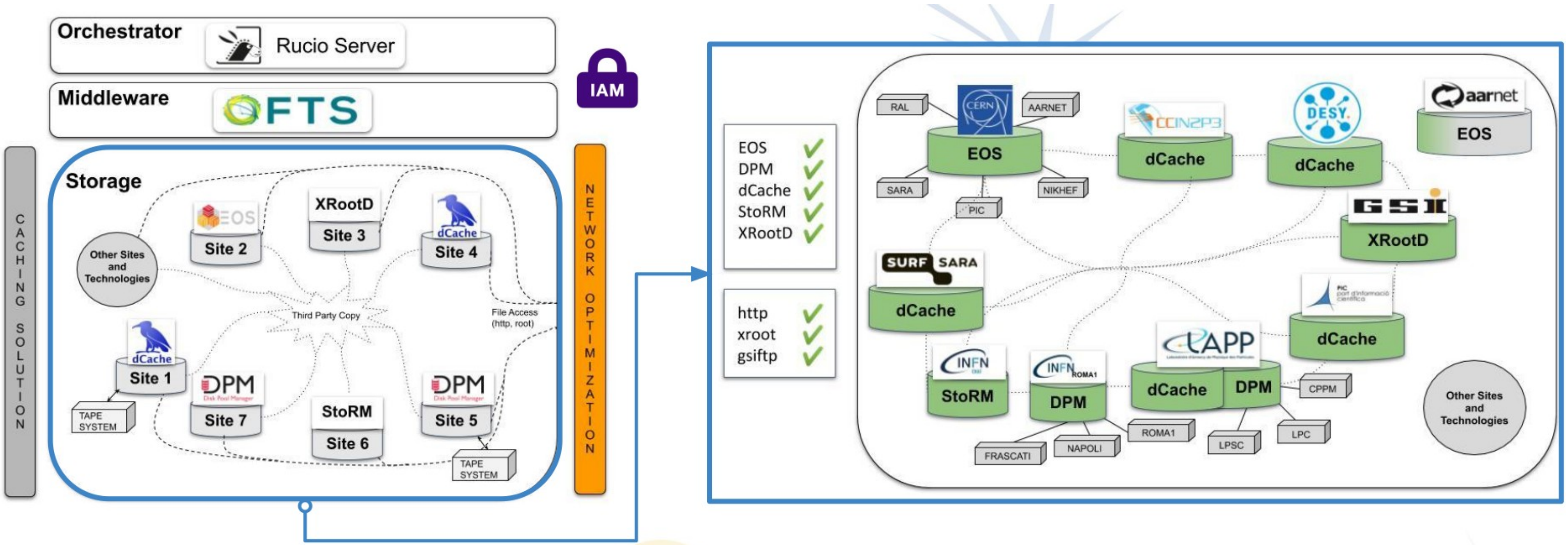
Syncronised services

- Accounting
- Service discovery
- VM images
- AAI
- DNS (*.desy.fedcloud.eu)

Source: wiki.egi.eu/wiki/Federated_Cloud_user_support
egi.eu/federation/egi-federated-cloud

The ESCAPE Data Lake

Hiding complexity and providing transparent access to data



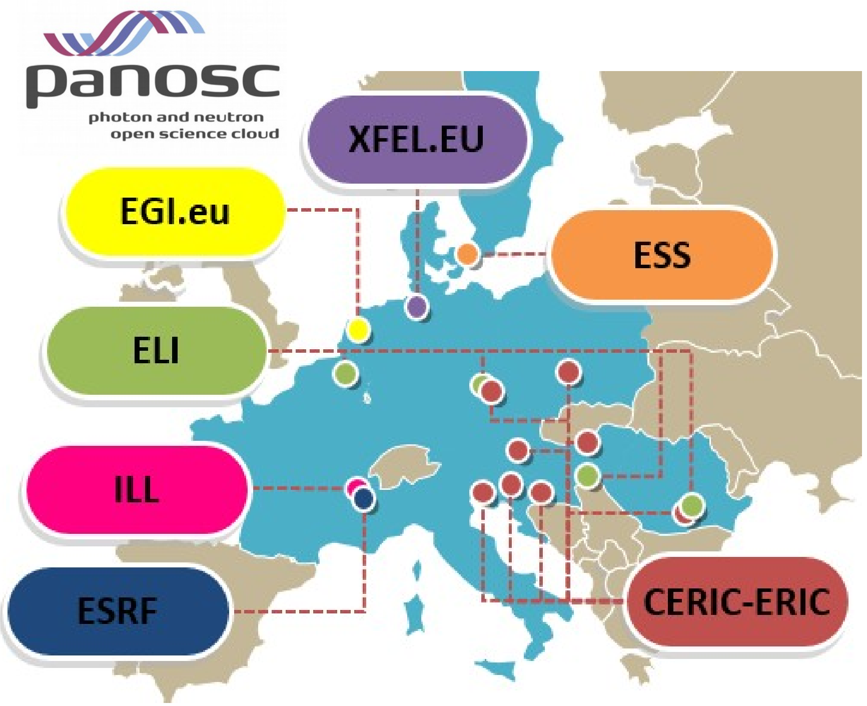
Heterogenous federated storage and operations model

Source: Slide by Xavier Espinal – PaN ESCAPE Data Management Workshop, 12 January 2012

ESCAPE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824064.

PaNOSC - Photon And Neutron Open Science Cloud

EOSC - European Open Science Cloud



EOSC

- FAIR data, effective Open Science

Networking | Compute | Storage | Sharing & Discovery

Data Management | Processing & Analysis | Security & Operations | Training & Support

Source: panosc.eu eosc-portal.eu

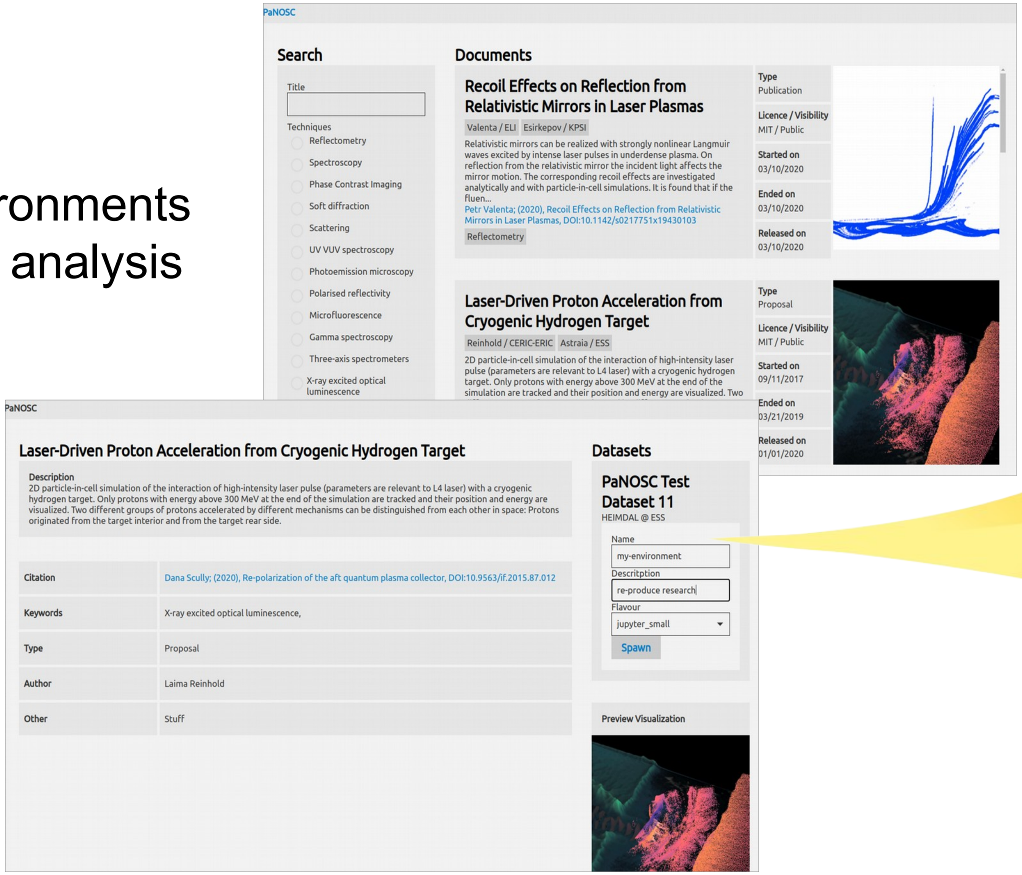
PaNOSC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852.

The PaNOSC/ExPaNDS (and others) use case

Interactive data analysis environments with Jupyter Notebooks

Science portals

- **F**ind data
- **A**ccess data
- **I**nteroperable environments
- **R**eproducible data analysis



PaNOSC

Search

Title

Techniques

○ Reflectometry
○ Spectroscopy
○ Phase Contrast Imaging
○ Soft diffraction
○ Scattering
○ UV VUV spectroscopy
○ Photoemission microscopy
○ Polarised reflectivity
○ Microfluorescence
○ Gamma spectroscopy
○ Three-axis spectrometers
○ X-ray excited optical luminescence

Documents

Recoil Effects on Reflection from Relativistic Mirrors in Laser Plasmas

Valenta / ELI · Esirkepov / KPSI

Relativistic mirrors can be realized with strongly nonlinear Langmuir waves excited by intense laser pulses in underdense plasma. On reflection from the relativistic mirror the incident light affects the mirror motion. The corresponding recoil effects are investigated analytically and with particle-in-cell simulations. It is found that if the fluen...

Petr Valenta; (2020), Recoil Effects on Reflection from Relativistic Mirrors in Laser Plasmas, DOI:10.1142/s0217751x19430103

Reflectometry

Type: Publication
Licence / Visibility: MIT / Public
Started on: 03/10/2020
Ended on: 03/10/2020
Released on: 03/10/2020

Laser-Driven Proton Acceleration from Cryogenic Hydrogen Target

Reinhold / CERIC-ERIC · Astraia / ESS

2D particle-in-cell simulation of the interaction of high-intensity laser pulse (parameters are relevant to L4 laser) with a cryogenic hydrogen target. Only protons with energy above 300 MeV at the end of the simulation are tracked and their position and energy are visualized. Two

Type: Proposal
Licence / Visibility: MIT / Public
Started on: 09/11/2017
Ended on: 03/21/2019
Released on: 01/01/2020

PaNOSC

Laser-Driven Proton Acceleration from Cryogenic Hydrogen Target

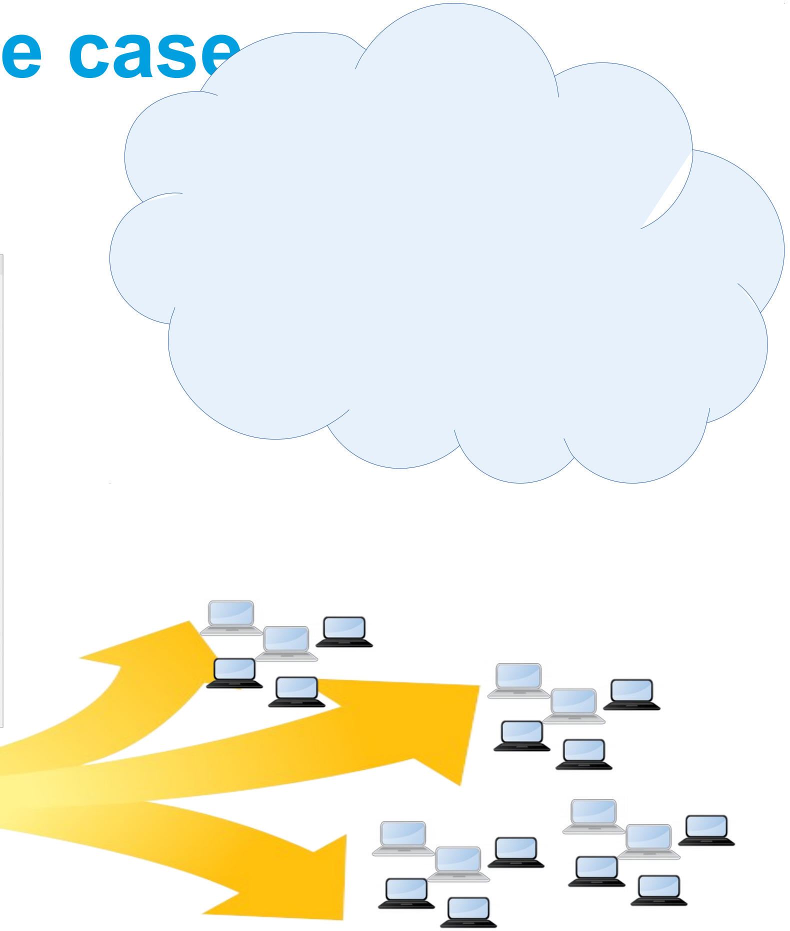
Description
2D particle-in-cell simulation of the interaction of high-intensity laser pulse (parameters are relevant to L4 laser) with a cryogenic hydrogen target. Only protons with energy above 300 MeV at the end of the simulation are tracked and their position and energy are visualized. Two different groups of protons accelerated by different mechanisms can be distinguished from each other in space: Protons originated from the target interior and from the target rear side.

Citation	Dana Scully; (2020), Re-polarization of the aft quantum plasma collector, DOI:10.9563/if.2015.87.012
Keywords	X-ray excited optical luminescence,
Type	Proposal
Author	Laima Reinhold
Other	Stuff

Datasets

PaNOSC Test Dataset 11
HEIMDAL @ ESS

Name: my-environment
Descritption: re-produce research
Flavour: jupyter_small



Source: github.com/panosc-portal

PaNOSC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852.

Software as a Service
Containerized applications
Deployments as code

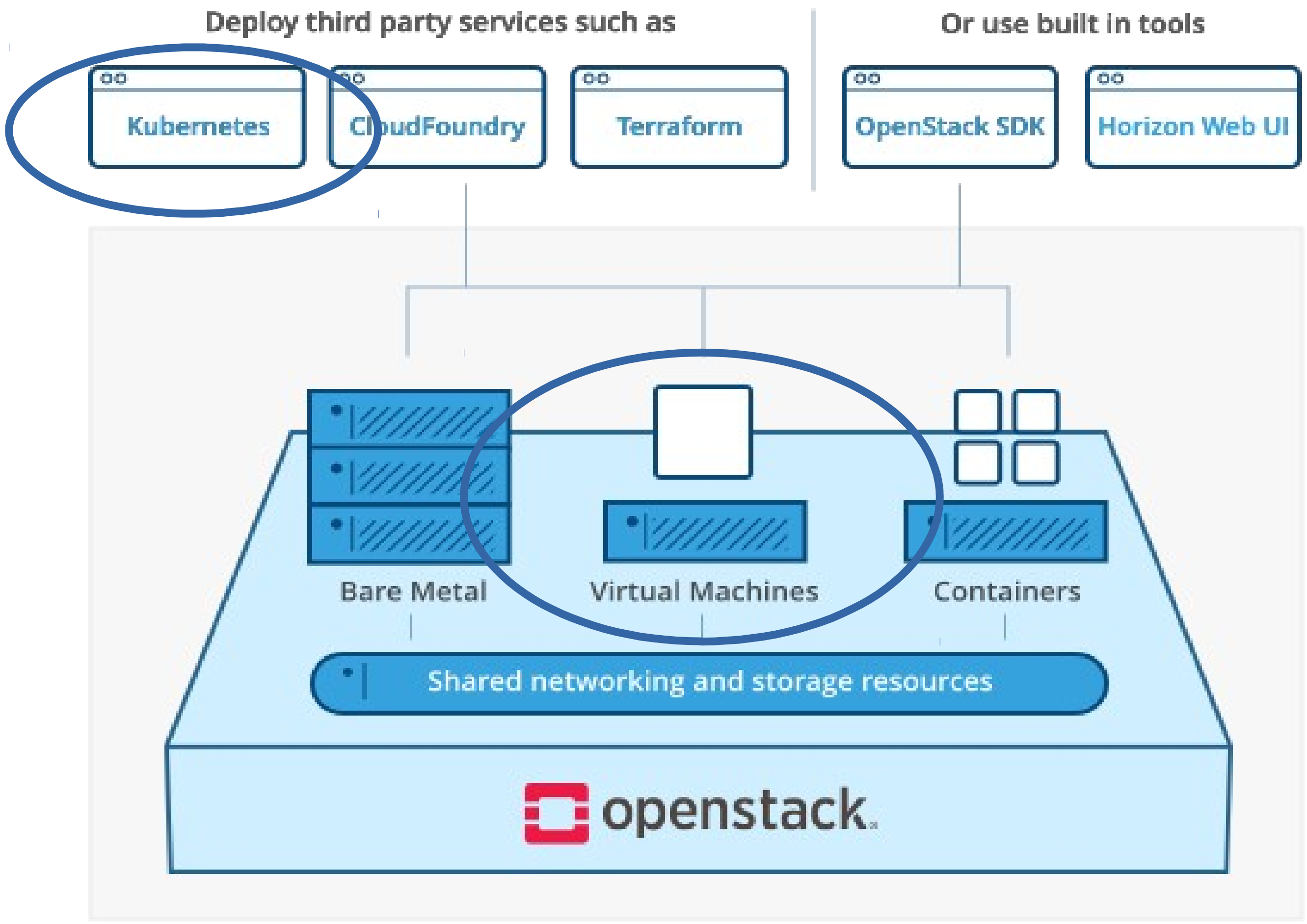
Orchestration
Rancher managed Kubernetes
Helm Package Manager

Containerization
Cloud Native CI/CD
Docker Registry

Infrastructure
Compute Cloud
Storage Systems

jupyterhub
DASK
Spark
Prometheus
Grafana
RANCHER
kubernetes
HELM
GitLab
openstack
ceph
dCache

Primary Field of Application in Cloud Computing: Container Orchestration with Kubernetes



Openstack

- Used as a **virtualization platform**

Kubernetes

- Clusters of **virtual machines**
- For **containerized applications,** automated deployments and scaling

Source: openstack.org/software

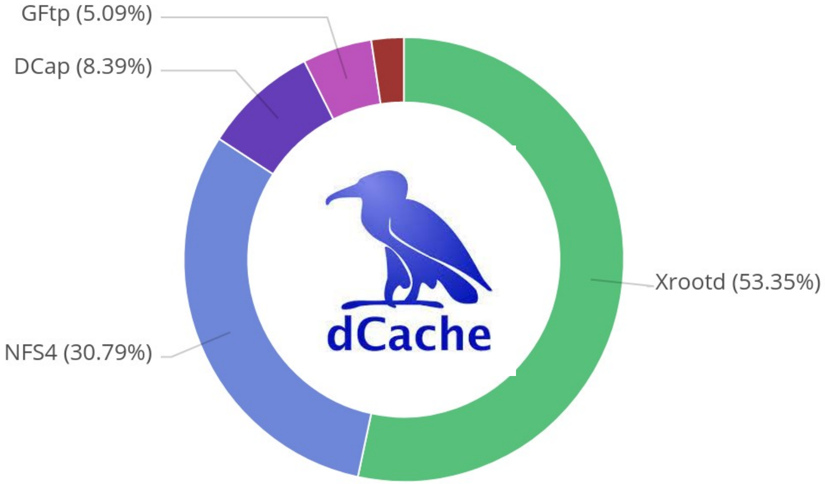
Storage in the Cloud

CEPH

- Block Storage for Openstack Cinder (RBD)
  - Disk storage attached to a virtual machine
  - Accessible from attached VM only
- Object Storage for Apps and Openstack Swift (S3)
  - MinIO S3 Proxy: Accessible from anywhere

dCache

- Shared file system
  - Disk storage attached to a virtual machine
  - Accessible from many VMs in parallel
- NFS, SMB
- Scale dcache-demo.desy.de to > 1PB
- **Storage for scientific data** (immutable)



dCache access stats at DESY

Software as a Service
Containerized applications
Deployments as code

Orchestration
Rancher managed Kubernetes
Helm Package Manager

Containerization
Cloud Native CI/CD
Docker Registry

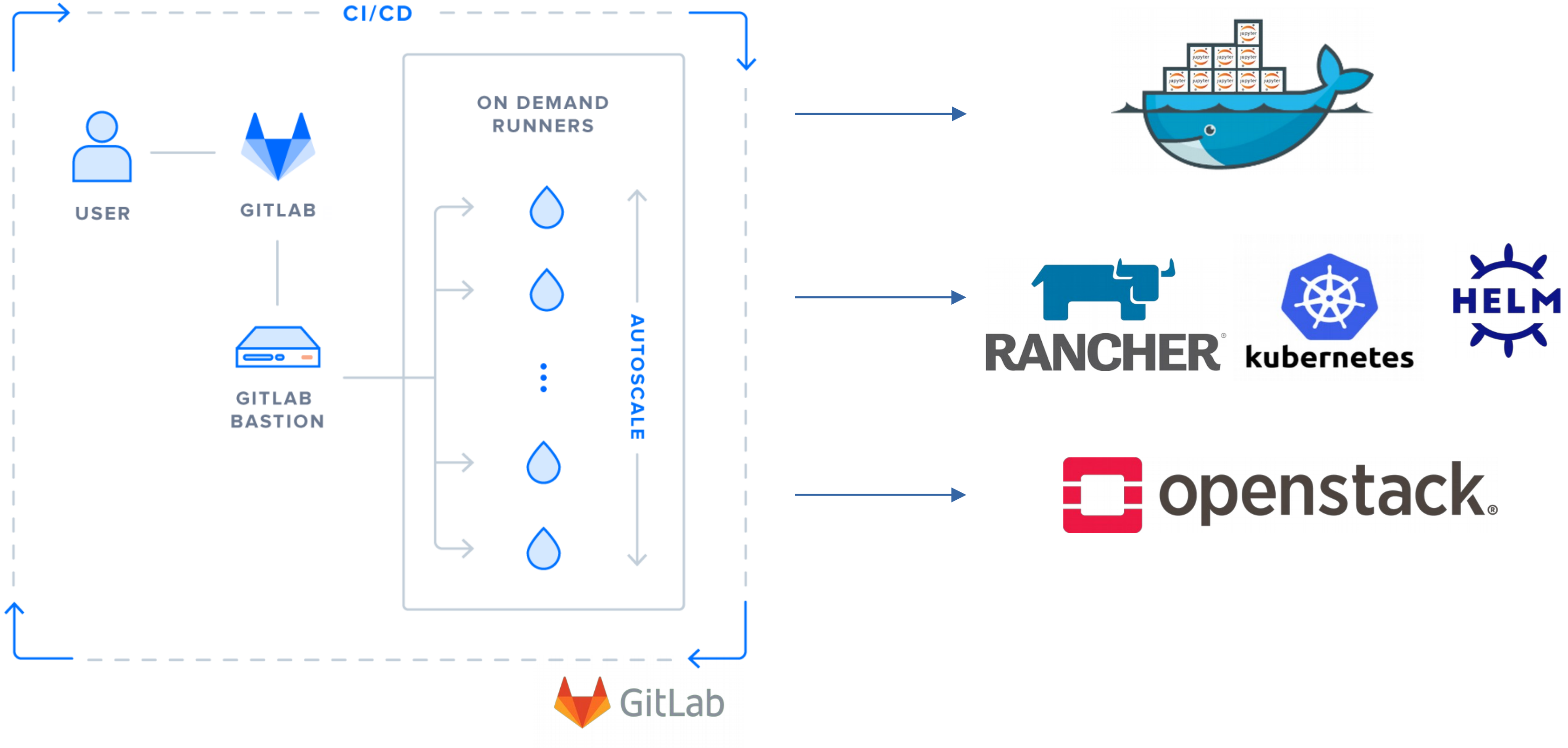
Infrastructure
Compute Cloud
Storage Systems

jupyterhub
DASK
Spark
Prometheus
Grafana
RANCHER®
kubernetes
HELM
GitLab
openstack®
ceph
dCache

GitLab CI/CD for Container and Cloud Applications

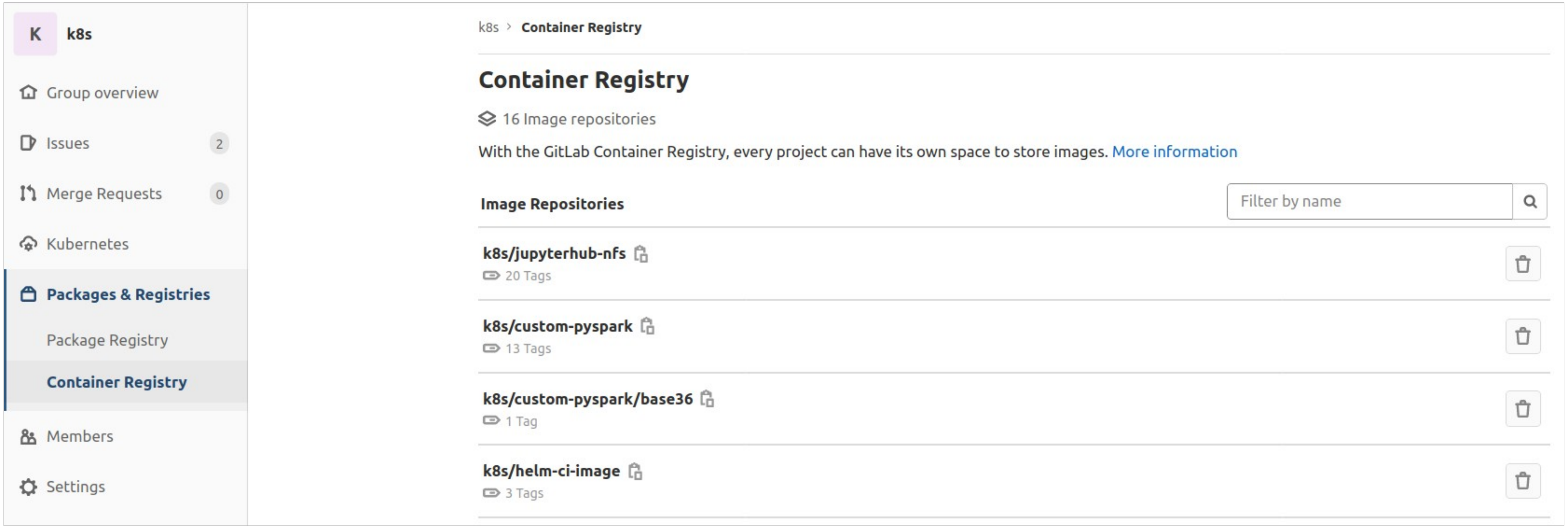
Git as a single source of truth for declarative infrastructure and application

- DevOps Platform
- Auto-scaling CI/CD
- Container Registry



Source: about.gitlab.com

Integrated Docker Registry in GitLab



Container Registry

16 Image repositories

With the GitLab Container Registry, every project can have its own space to store images. More information

Image Repositories

k8s/jupyterhub-nfs
20 Tags

k8s/custom-pyspark
13 Tags

k8s/custom-pyspark/base36
1 Tag

k8s/helm-ci-image
3 Tags

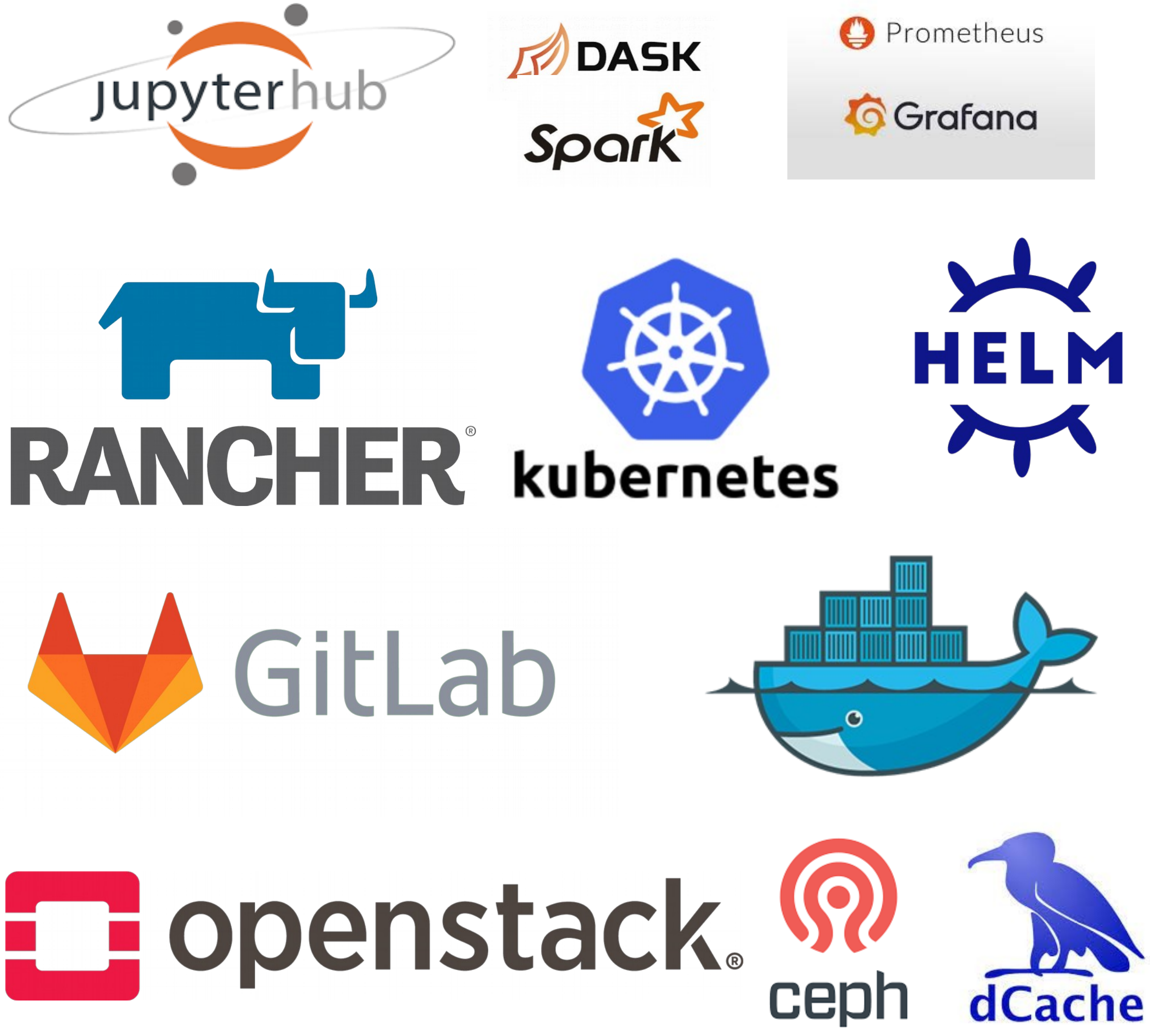
- Host public and private Container Images
  - Docker (Container Registry)
  - Singularity
    - As Docker Image
    - Singularity Images as build artifacts

Software as a Service
Containerized applications
Deployments as code

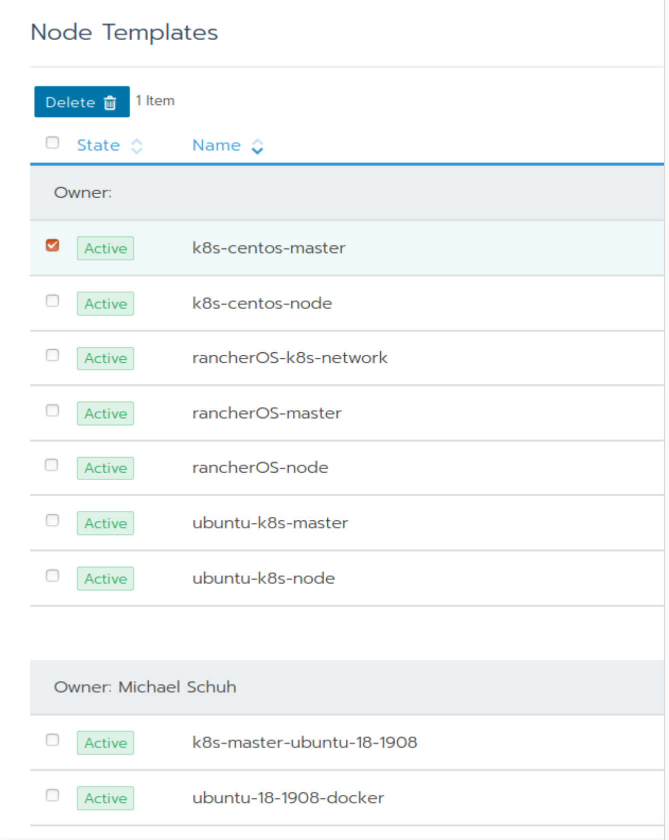
Orchestration
Rancher managed Kubernetes
Helm Package Manager

Containerization
Cloud Native CI/CD
Docker Registry

Infrastructure
Compute Cloud
Storage Systems



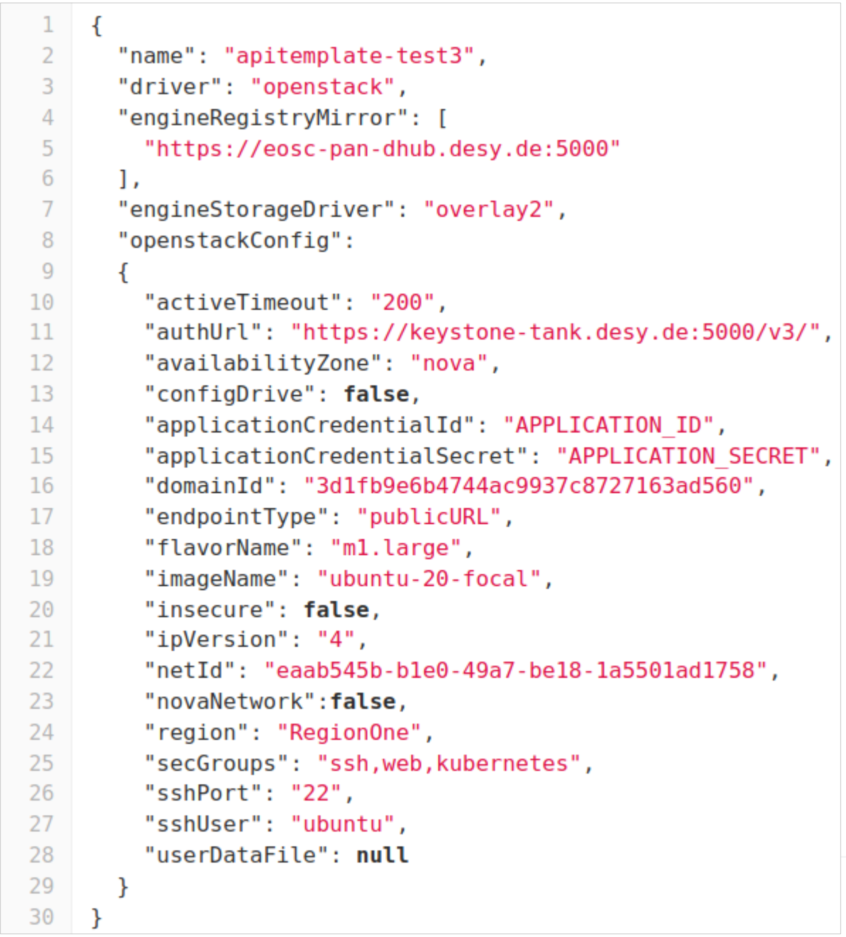
Rancher Node Templates



Node Templates

Delete 🗑 1 Item

□	State	Name
	Owner:	
☑	Active	k8s-centos-master
□	Active	k8s-centos-node
□	Active	rancherOS-k8s-network
□	Active	rancherOS-master
□	Active	rancherOS-node
□	Active	ubuntu-k8s-master
□	Active	ubuntu-k8s-node
	Owner: Michael Schuh	
□	Active	k8s-master-ubuntu-18-1908
□	Active	ubuntu-18-1908-docker



```json
{
  "name": "apitemplate-test3",
  "driver": "openstack",
  "engineRegistryMirror": [
    "https://eosc-pan-dhub.desy.de:5000"
  ],
  "engineStorageDriver": "overlay2",
  "openstackConfig":
  {
    "activeTimeout": "200",
    "authUrl": "https://keystone-tank.desy.de:5000/v3/",
    "availabilityZone": "nova",
    "configDrive": false,
    "applicationCredentialId": "APPLICATION_ID",
    "applicationCredentialSecret": "APPLICATION_SECRET",
    "domainId": "3d1fb9e6b4744ac9937c8727163ad560",
    "endpointType": "publicURL",
    "flavorName": "m1.large",
    "imageName": "ubuntu-20-focal",
    "insecure": false,
    "ipVersion": "4",
    "netId": "eaab545b-b1e0-49a7-be18-1a5501ad1758",
    "novaNetwork":false,
    "region": "RegionOne",
    "secGroups": "ssh,web,kubernetes",
    "sshPort": "22",
    "sshUser": "ubuntu",
    "userDataFile": null
  }
}
```

- Openstack VMs as k8s nodes
- Node pools
  - Workers
  - Control Plane

Additional software components

“bare” Kubernetes is not enough

Nginx Ingress Controller

- Direct traffic to pods

MetalLB Loadbalancer

- Level2 Loadbalancer for Kubernetes

Cinder Storage Class

- Automatically Provision Volumes in Ceph

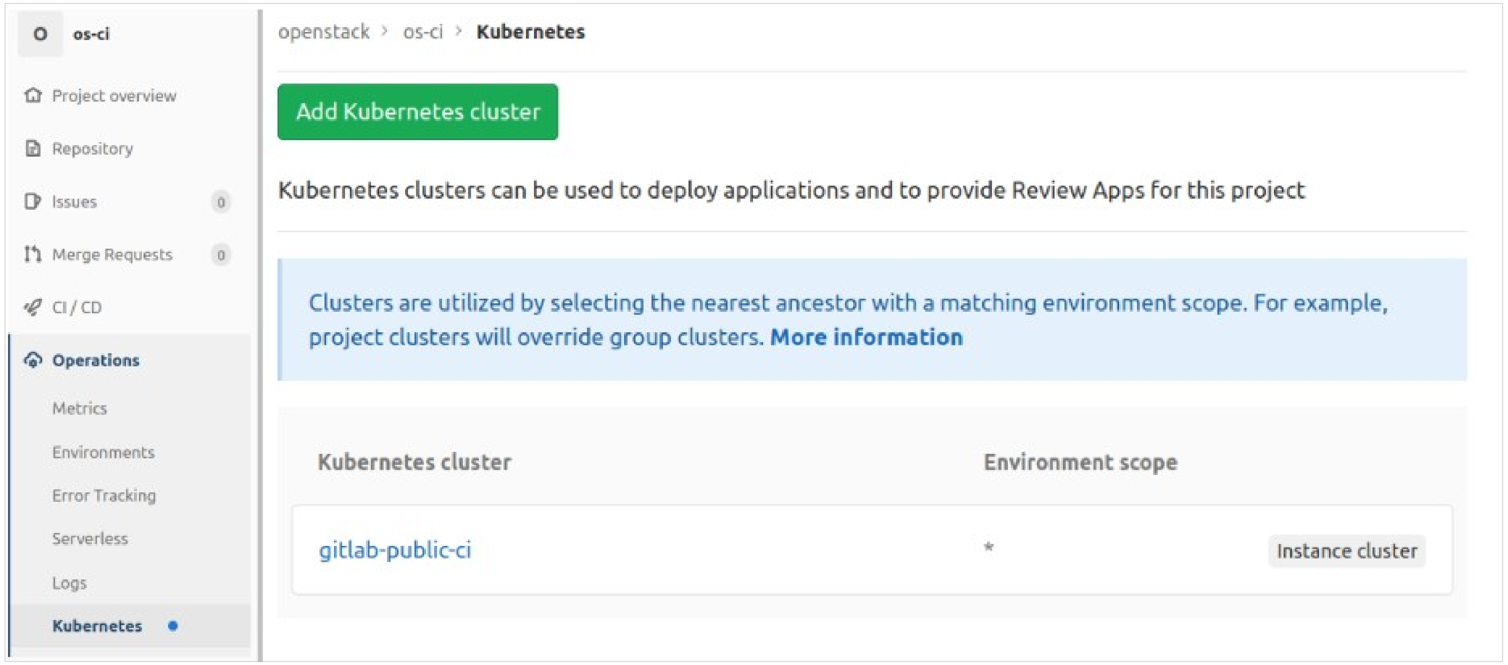
Cert Manager

- Provides Let’s Encrypt Certificates
- Watches the Kubernetes API for *Ingress* Objects

dCache

- Shared Filesystem on cluster nodes

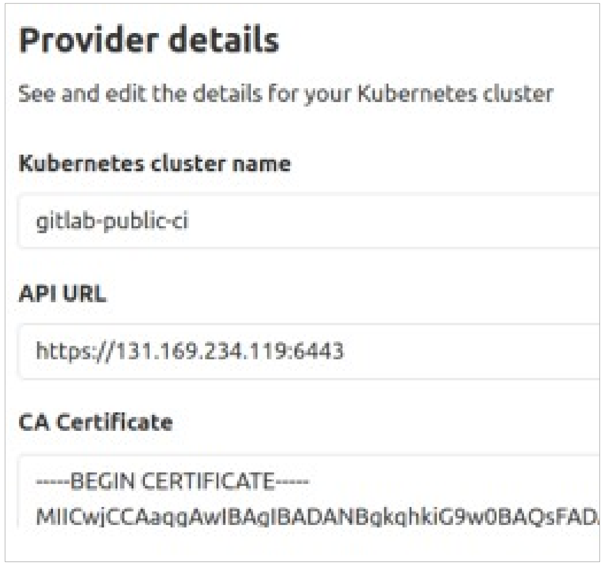
Integrate Kubernetes with Gitlab



Kubernetes clusters can be used to deploy applications and to provide Review Apps for this project

Clusters are utilized by selecting the nearest ancestor with a matching environment scope. For example, project clusters will override group clusters. **More information**

Kubernetes cluster	Environment scope	
gitlab-public-ci	*	Instance cluster



Provider details

See and edit the details for your Kubernetes cluster

Kubernetes cluster name

gitlab-public-ci

API URL

https://131.169.234.119:6443

CA Certificate

```
-----BEGIN CERTIFICATE-----
MIICwjCCAaqgAwIBAgIBADANBgkqhkiG9w0BAQsFAD
```

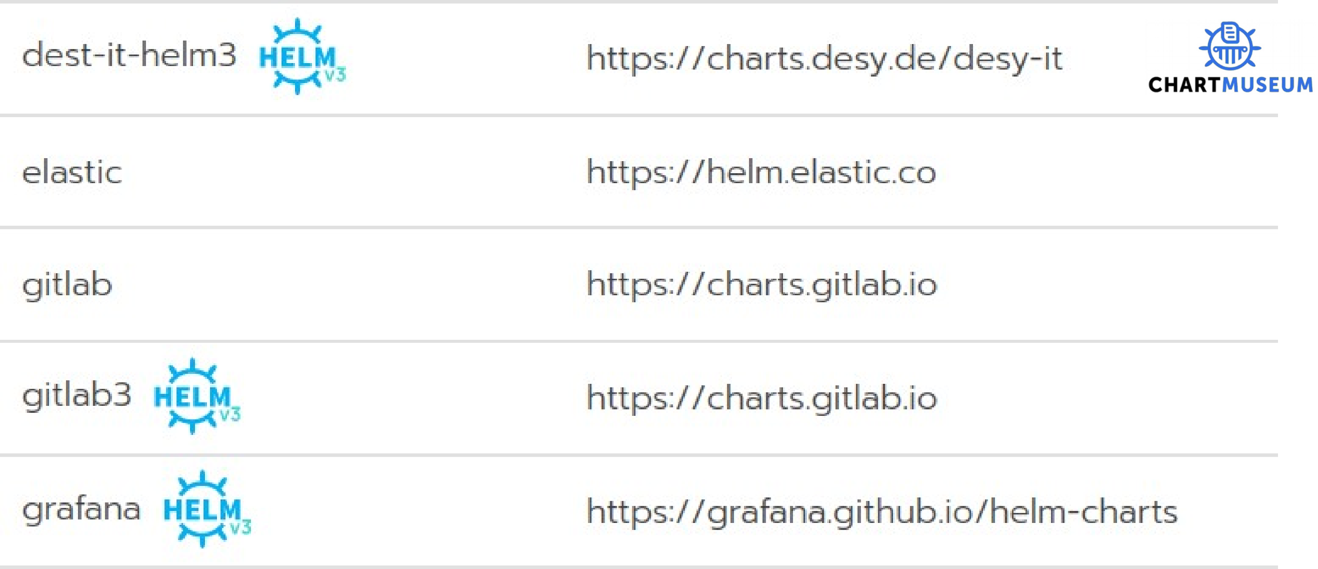
Deliver Kubernetes as a Service for GitLab users

- Instance cluster
- Group clusters
- Project clusters

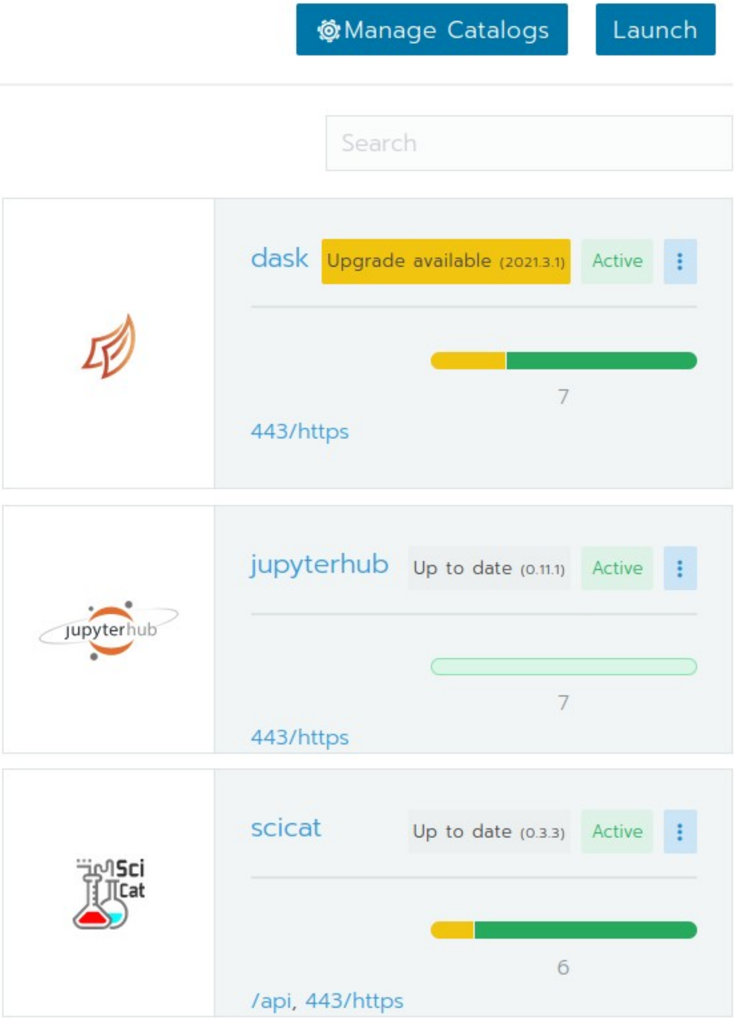
- Users deploy environments for review and production

Managing Apps with Helm

Chart Repository



dest-it-helm3 HELM v3	https://charts.desy.de/desy-it CHARTMUSEUM
elastic	https://helm.elastic.co
gitlab	https://charts.gitlab.io
gitlab3 HELM v3	https://charts.gitlab.io
grafana HELM v3	https://grafana.github.io/helm-charts



dask Upgrade available (2021.3.1) Active
7
443/https

jupyterhub Up to date (0.11.1) Active
7
443/https

scicat Up to date (0.3.3) Active
6
/api, 443/https

charts.desy.de

- Templated k8s definition files
- Repository for Helm Chart Tarballs
  - Add as Rancher Catalog
- Install charts as Rancher Apps

Software as a Service
Containerized applications
Deployments as code

Orchestration
Rancher managed Kubernetes
Helm Package Manager

Containerization
Cloud Native CI/CD
Docker Registry

Infrastructure
Compute Cloud
Storage Systems

jupyterhub
DASK
Spark
Prometheus
Grafana
RANCHER
kubernetes
HELM
GitLab
openstack
ceph
dCache

Managing Apps with Helm

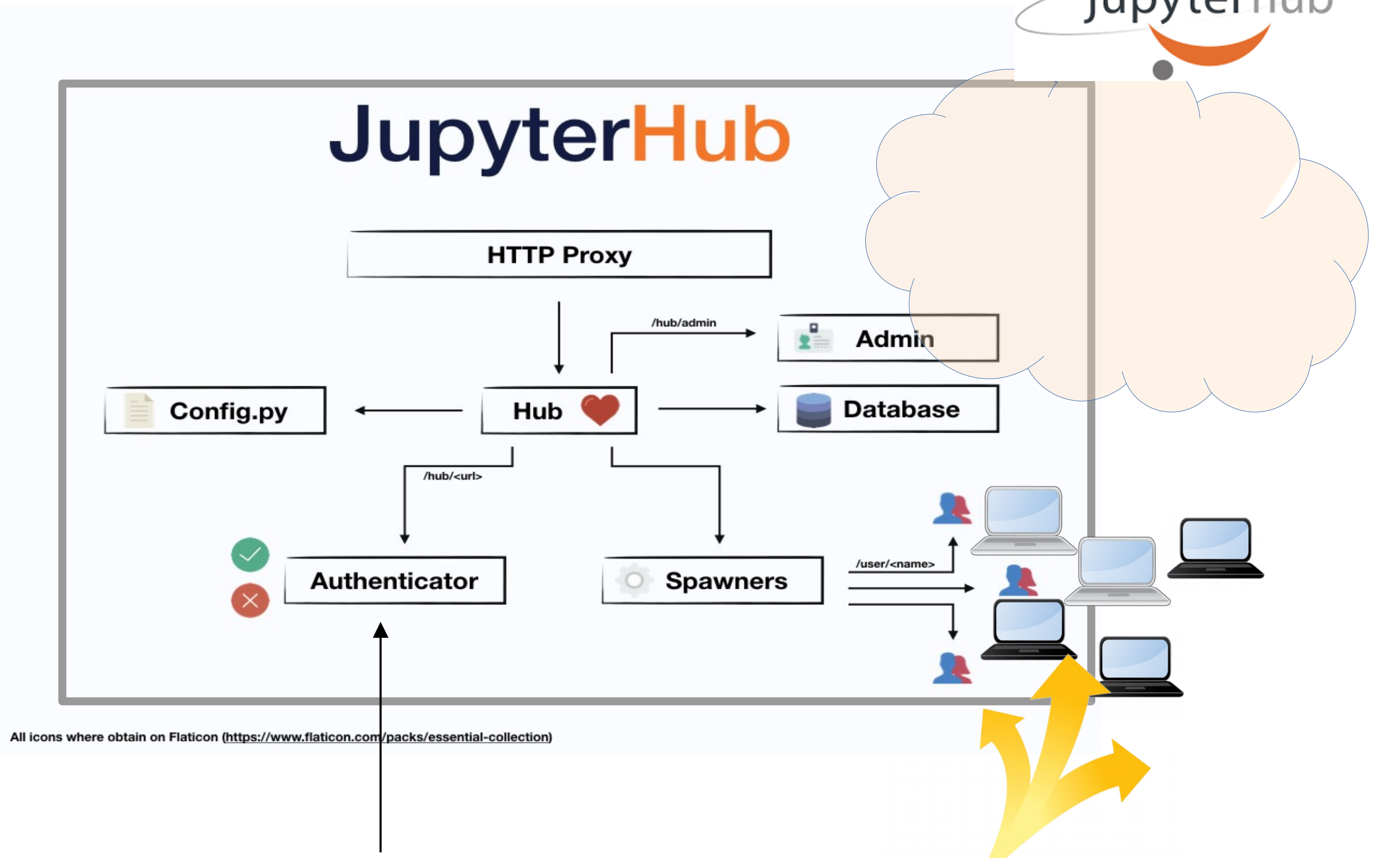
Deployment of Jupyterhub

Deployment

- Add DNS alias to Loadbalancer IP
- Add Helm Chart repository
- Customize Values.yaml
- Install to k8s (helm install)

Map Role Based Group Memberships on OIDC Proxy to local accounts and UID/GIDs

- Merge group memberships from user federation in Keycloak
- Export UID/GIDs as non-standard attribute in OIDC Token
- Run Jupyter Servers with UID/GIDs



All icons where obtain on Flaticon (https://www.flaticon.com/packs/essential-collection)

KEYCLOAK

User Federation

- LDAP (DESY)
- OIDC (EGI Check-in)
- OIDC (Helmholtz AAI)

dCache

PetaByte Storage

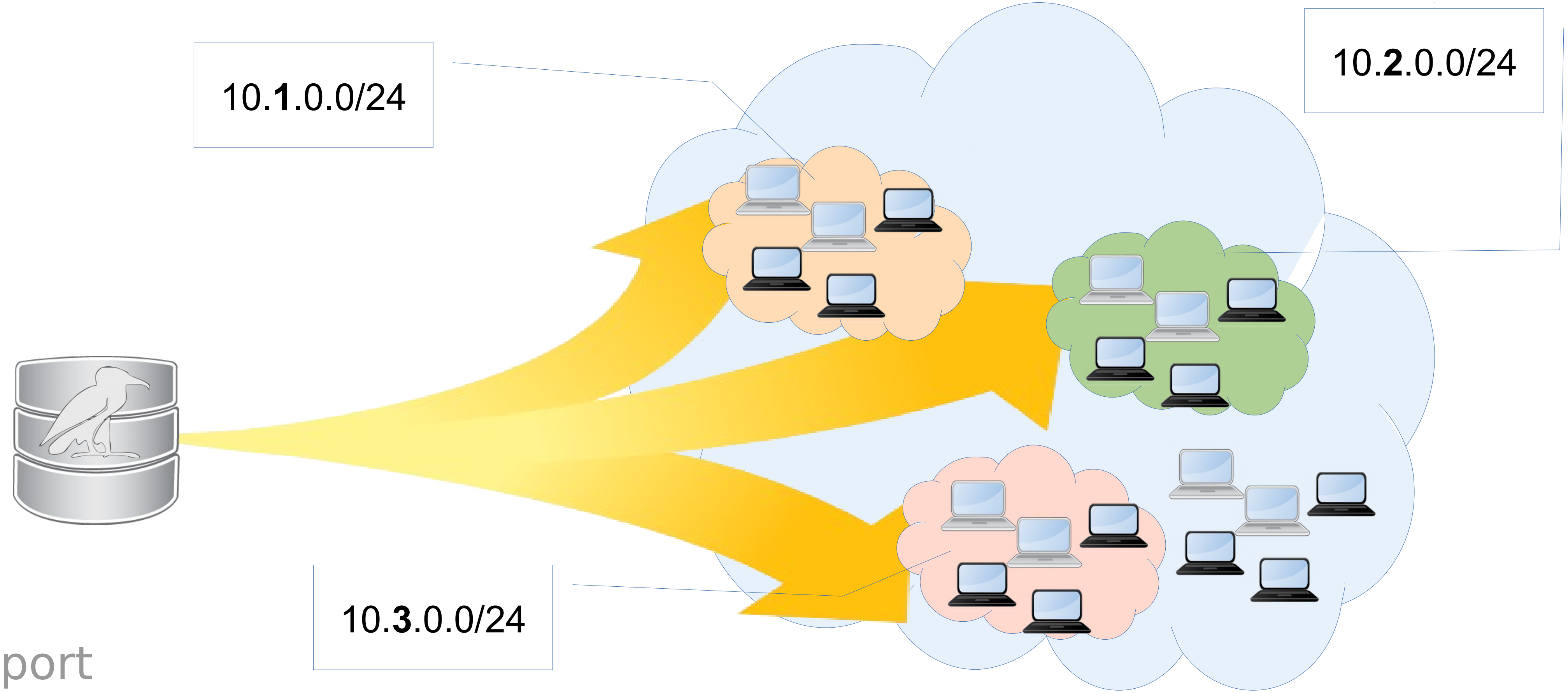
- NFS Mounts

NFS in elastic cloud environments

- Storage system can’t trust to OS level authentication/mapping
  - Users build and select images VM and Container images
- Storage system can’t trust client’s IP address
  - Use of public networks
  - After disposal VM’s IP returns to shared pool

- **NFSv3** based on trusted hosts
  - Server exports based on IP address
  - OS is responsible for proper mapping
- **NFSv4.0+**
  - Strong authentication is enforced
    - Krb5 + LDAP/AD
      - No kerberos infrastructure provided by public clouds
  - Backward compatibility is agreed for migration period
- **On the field, most of sites run NFSv4.0+ in NFSv3 security mode**

Map VM by IP or subnet to a dCache user



```
# /etc/export

/data 10.1.0.0/24(rw,all_squash,anonuid=1001,anongid=1001)
/data 10.2.0.0/24(rw,all_squash,anonuid=1002,anongid=1002)
```

dCache REST interface

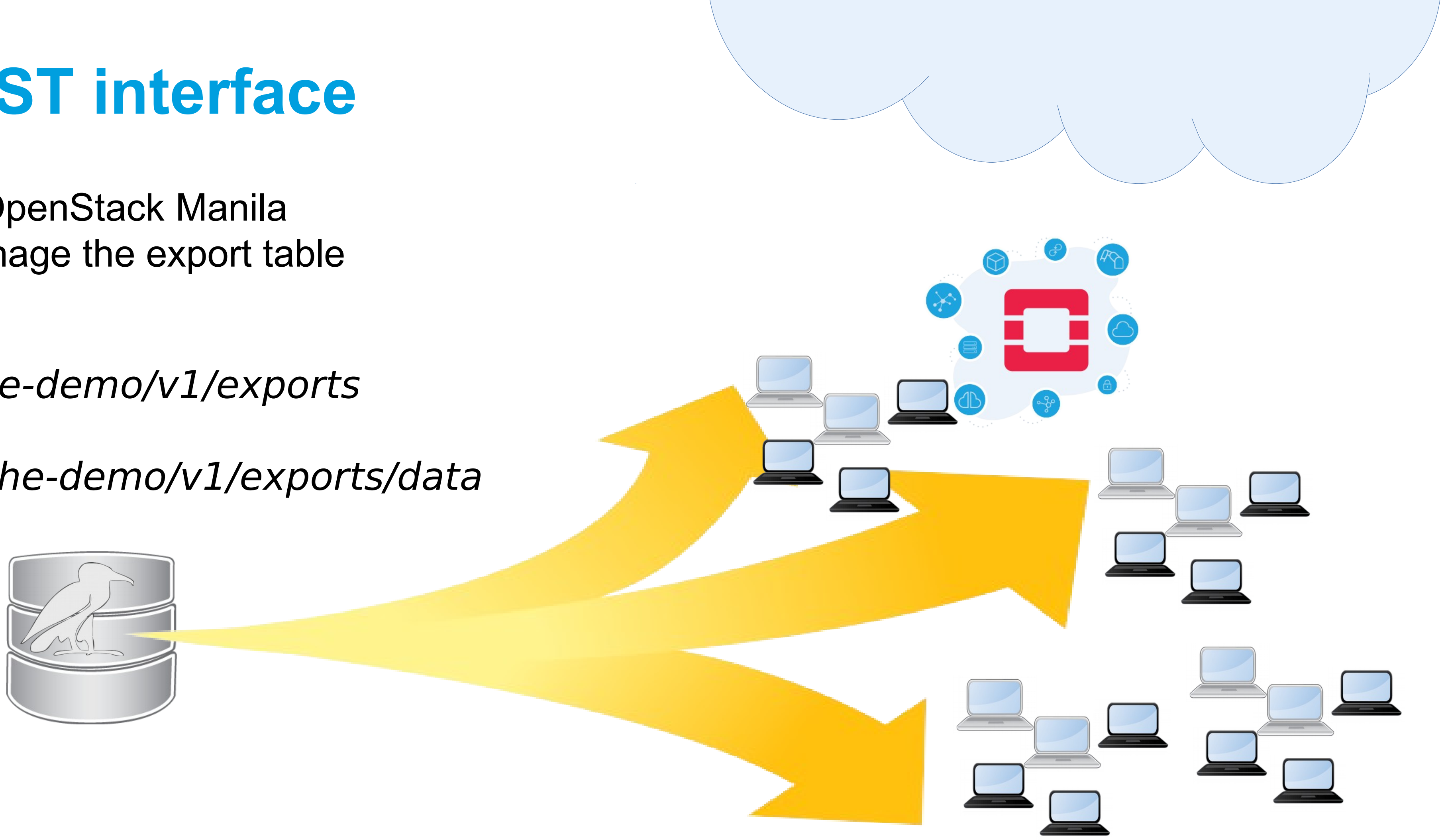
- Compatible with OpenStack Manila
- Simple API to manage the export table

Get defined exports

GET https://dcache-demo/v1/exports

Create share '/data'

POST https://dcache-demo/v1/exports/data



Summary and outlook

- dCache developers work on better cloud integration
  - Manage shared storage with exports REST API
  - Geo-aware zones
- NFS community works to address cloud challenges
  - RPC-over-TLS
  - 3rd party copy
- Jupyter Hub extensions
  - for ESCAPE Data Lake
  - for Remote Desktops

The 15th International dCache workshop 2021 will take place from 2021-06-01 to 2021-06-02 as a virtual event.
indico.desy.de/event/29564

Contact

DESY.
Deutsches Elektronen-Synchrotron
www.desy.de

Michael Schuh
Research and Innovation
michael.schuh@desy.de

Tigran Mkrtchyan
dcache.org
tigran.mkrtchyan@desy.de