Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

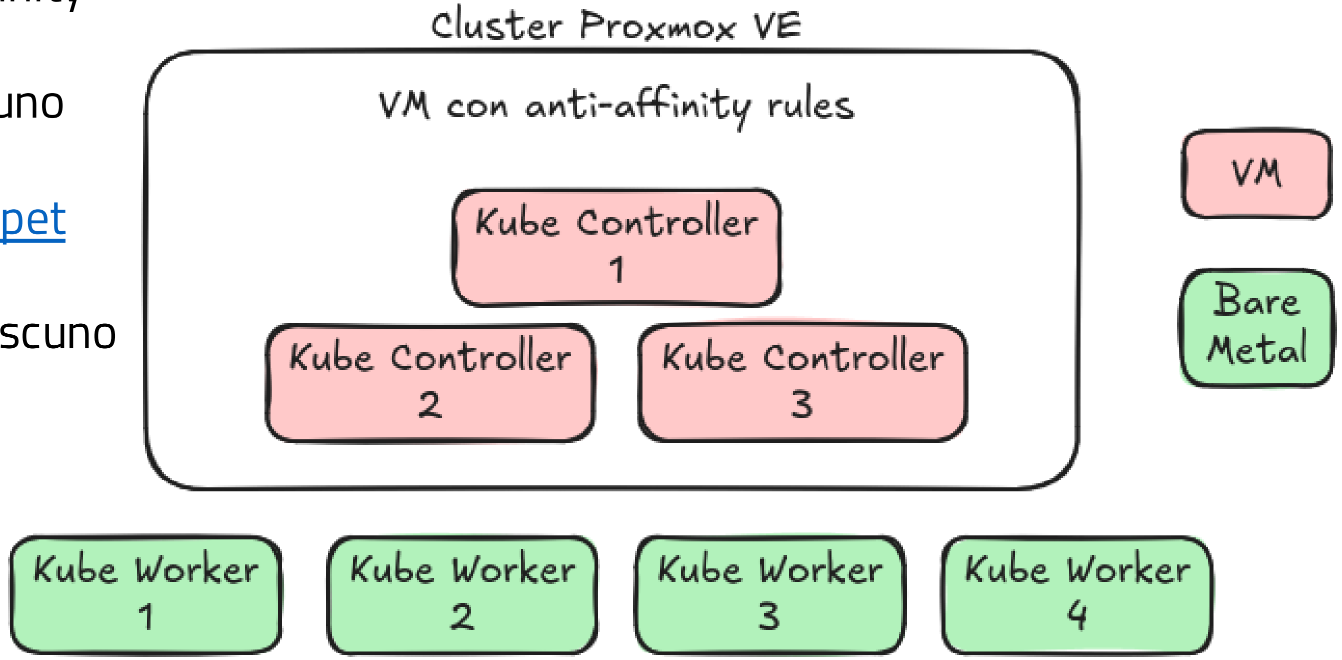
Cluster Kubernetes RKE2 con Capsule, Ceph e OIDC – Progetto ICSC

Implementazione HA, multi-tenancy e autenticazione IAM INFN

Meeting annuale ATLAS Italia Computing 2025, CNAF – Bologna

Architettura del Cluster Kubernetes

- 3 nodi controller come VM su Proxmox VE in anti-affinity
- 4 nodi worker bare metal: 96 core, 500Gi Mem ciascuno
- Installazione Kubernetes (RKE2) tramite modulo Puppet
- In futuro altri 8 nodi worker: 192 core, 1.5Ti Mem ciascuno



HA e Storage

Configurazione HA, load balancing e monitoring:

- Kube-VIP configura e gestisce un Virtual IP condiviso tra i nodi controller
- Garantisce l'accesso continuo all'API server anche in caso di failover
- Deployment via DaemonSet sui nodi controller
- Monitoring con Prometheus e Grafana

Kube-VIP

Configurazione storage:

- Storage Ceph in produzione al Tier 2 Roma
- Nuovo hardware appena arrivato per creare un cluster Ceph ICSC dedicato
- StorageClass di default configurata con ceph-rbd-csi
- Provisioning dinamico grazie al CSI plugin
- Pool dedicato al cluster Kubernetes ICSC

Capsule per Multi-Tenancy

Kubernetes non gestisce nativamente utenti/gruppi come “tenants”:

- Isolamento manuale tramite namespace + RBAC
- Configurazioni complesse e soggette ad errori

Capsule:

- Introduce un layer di astrazione chiamato Tenant
- Automatizza isolamento, policy, quote e controllo degli accessi

Deployment rapido via [helm chart] e [documentazione] chiara e completa

```yaml
apiVersion: capsule.clastix.io/v1beta2
kind: Tenant
metadata:
  name: tenant-example
spec:
  owners:
    - kind: Group
      name: admins/beta-testers
  storageClasses:
    default: csi-rbd-sc
    allowed:
    - csi-rbd-sc
  namespaceOptions:
    quota: 5
  resourceQuotas:
    scope: Tenant
    items:
    - hard:
        limits.cpu: "8"
        limits.memory: 16Gi
        requests.cpu: "8"
        requests.memory: 16Gi
        pods: "10"
        services: "15"
        secrets: "5"
        requests.storage: 200Gi
        persistentvolumeclaims: "10"
```

Autenticazione & Autorizzazione con OIDC e Capsule

- Configurazione lato cluster Kubernetes (RKE2)
- Autenticazione integrata con INFN IAM via OIDC (JWT Token)
- Autorizzazione basata su gruppi e tenants con Capsule
- Accesso granulare e limitato alle risorse secondo policy assegnate

Sui controller: /etc/rancher/rke2/config.yaml

```
kube-apiserver-arg:
 - --oidc-issuer-url=https://iam.cloud.infn.it/ # OIDC issuer URL
 - --oidc-client-id=icsc-k8s-roma1 # Must match the 'aud' claim in the JWT
 - --oidc-username-claim=preferred_username # Claim containing the username
 - --oidc-groups-claim=groups # Claim containing the groups
 - --oidc-username-prefix=- # Disables username prefix
```

Esempio di utilizzo

```
$ eval `oidc-agent-service use`
$ oidc-gen <username> --iss https://iam.cloud.infn.it -w device # Solo la prima volta
$ oidc-add <username>
$ token=$(oidc-token --aud icsc-k8s-roma1 <username>)
$ kubectl --token=$token <kubectl command>
```

Conclusioni

- Cluster attualmente operativo
- Utilizzabile via CLI previa creazione di tenant
- In corso test da WP5 INFN DataCloud per integrazione nella dashboard

Documentazione installazione e configurazione su INFN Confluence:

- Authentication and Authorization Procedure on Kubernetes with JWT Generated via oidc-agent
- Configuring Kube-VIP for High Availability in Kubernetes RKE2
- K8S - monitoring (Prometheus + Grafana)