Cloud GARR: L'esperienza d’uso in D4Science

Andrea Dell’Amico
CNR-ISTI
D4Science.org

VREs (Virtual Research Environments) support research projects and open science

- Managed services (JupyterHub, Rstudio, Shinyproxy, CKAN-based catalogues, Geoserver, Geonetwork, Hadoop, etc.)
- (mostly) unified storage
- Authentication and authorization
- HA and redundancy wherever is possible
D4Science gateway for OpenAIRE: https://tools.openaire.eu
Use case: Protected Area Impact Maps
https://i-marine.d4science.org/web/protectedareaimpactmaps

Services and resources:

- (Authentication)
- (Authorization)
- Static web Application
- Analytics Engine (Dataminer)
- Geoserver
- Workspace
- FAIR
Deployment: everything is automated (mostly)

Ansible is the provisioning tool used to configure all the servers and all the services.

- All the servers share a basic configuration set (language, ssh access, timezone, firewall, dns resolver, NTP, ...)
- No manual editing of configuration files
- A new instance of a server can be ready in a matter of minutes
D4Science – GARR: schema
GARR – D4Science: how it started

• One Region (GARR-CT1) to test the service
• Some dev service
• Provisioning of the Vms
• Instances activated manually from the dashboard
GARR – D4Science: now

- Three regions

- Production services
  - Rstudio
  - Analytics Engine
  - Hadoop Cluster
  - Elasticsearch Clusters
  - JupyterHub
  - Kubernetes
  - ...

- Instances managed by ansible exploiting the OpenStack API
GARR – D4Science: how we use the infras

- @GARR: 135 VM, 20+ TB of disk, 3.3 TB of RAM
- @CNR: 463 VM, 800 TB of disk, 6+ TB of RAM
- In the GARR regions:
  - Bigger instances
  - Instances have a shorter lifespan
  - Cluster of services to ensure reliability (analytics engine, JupyterHub, Rstudio)
- We still have SPoF (some services cannot have remote replicas, DNS shortcomings, etc.)
GARR – D4Science: how it looks
GARR cloud: what gives us

- Quite a lot of resources
- Reliability
- The opportunity of offloading some workload when we need to
- APIs
GARR cloud: what we would like to have

- Better storage options: true S3, remote Posix FS (*Data Lakes* is the buzzword of the second half of 2021)
- Availability Zones
- VPN as a service
- Load balancer as a service
GARR cloud future: federation?

- Give to the federation a subset of the internal resources?
- OpenStack constraints?
  - Software version
  - Installation method?
- What (legal) terms?
- Other?
Federated cloud: opportunities!

- Build competences: cloud, distributed storage, monitoring and observability of complex infrastructures.
- Develop solutions tailored to the research community
- Avoid the vendor lock-ins (every commercial cloud provider has its own services, with their APIs and functionalities. Moving out of them is challenging and expensive)