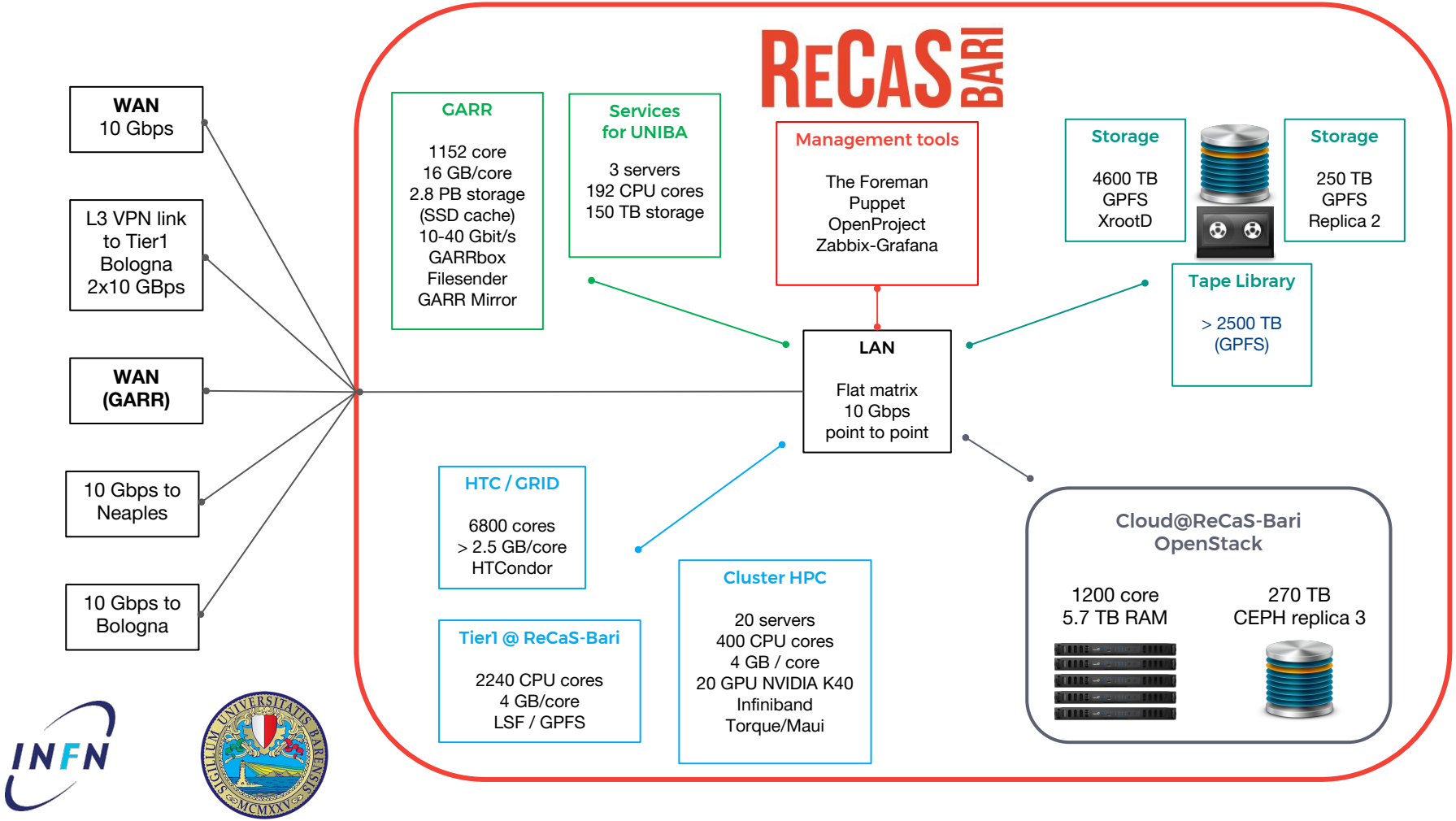
INFN Technological Infrastructure for biological data analysis

Stefano Nicotri

INFN - Istituto Nazionale di Fisica Nucleare - Sezione di Bari, Italy

Workshop on Computational Metagenomics: Methods, Standards and Experimental Procedures
June 19-20, 2017 - University of Bari "Aldo Moro", Bari, Italy


ReCaS BARI
WAN
10 Gbps
L3 VPN link
to Tier1
Bologna
2x10 GBps
WAN
(GARR)
10 Gbps to
Neaples
10 Gbps to
Bologna
GARR
1152 core
16 GB/core
2.8 PB storage
(SSD cache)
10-40 Gbit/s
GARRbox
Filesender
GARR Mirror
Services
for UNIBA
3 servers
192 CPU cores
150 TB storage
Management tools
The Foreman
Puppet
OpenProject
Zabbix-Grafana
Storage
4600 TB
GPFS
XrootD
Storage
250 TB
GPFS
Replica 2
Tape Library
> 2500 TB
(GPFS)
LAN
Flat matrix
10 Gbps
point to point
HTC / GRID
6800 cores
> 2.5 GB/core
HTCondor
Tier1 @ ReCaS-Bari
2240 CPU cores
4 GB/core
LSF / GPFS
Cluster HPC
20 servers
400 CPU cores
4 GB / core
20 GPU NVIDIA K40
Infiniband
Torque/Maui
Cloud@ReCaS-Bari
OpenStack
1200 core
5.7 TB RAM
270 TB
CEPH replica 3

ReCaS-Bari Cloud Infrastructure

The IaaS (Infrastructure as a Service) cloud platform **Cloud@ReCaS-Bari**, hosted at the **ReCaS-Bari** data center, provides **infrastructural computing resources** following the cloud computing paradigm.

Cloud@ReCaS-Bari

Total cores: ~ 1300 physical

Total RAM: ~ 5.7 TB

CMF: Openstack (Mitaka)

Storage:

- Ceph: 270 TB (replica 3). Pools: images, VMs, volumes, backups
- Swift (Object Storage): ~ 24 TB with replica 3

Network: VLAN setup with linuxbridge (no overlay network)

Compute: KVM hypervisors with RBD support

The IaaS cloud platform @ INFN Bari / UNIBA

- Resources (instances, or virtual machines, VM) can be used to develop and deploy software systems;

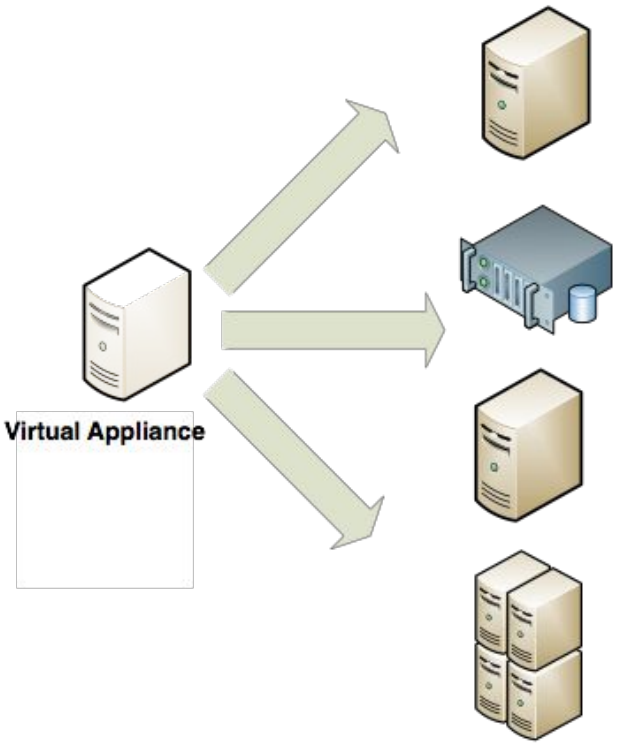
- It is possible to create **resilient systems** with high-availability using multiple instances (together with services provided by the IaaS infrastructure, as load-balancing and auto-scaling)

- Virtual instances are very similar to traditional hardware servers:
  - They use familiar Operating Systems (OS), as Linux, Windows, etc.
  - Any software compatible with the OS can be executed on them
  - Associating a public IP to the VM it is possible to interact with it through standard methods (ssh, RDP,…)

Image Service and Marketplace

- Pre-configured virtual images (*templates*) can be used to create virtual machines of different kinds (*flavor*) depending on the RAM and CPU required by your application.

- A certain number of templates (software configuration) is already available from the catalog, but the user can upload her/his own (also starting from *snapshots* of her/his own VMs).



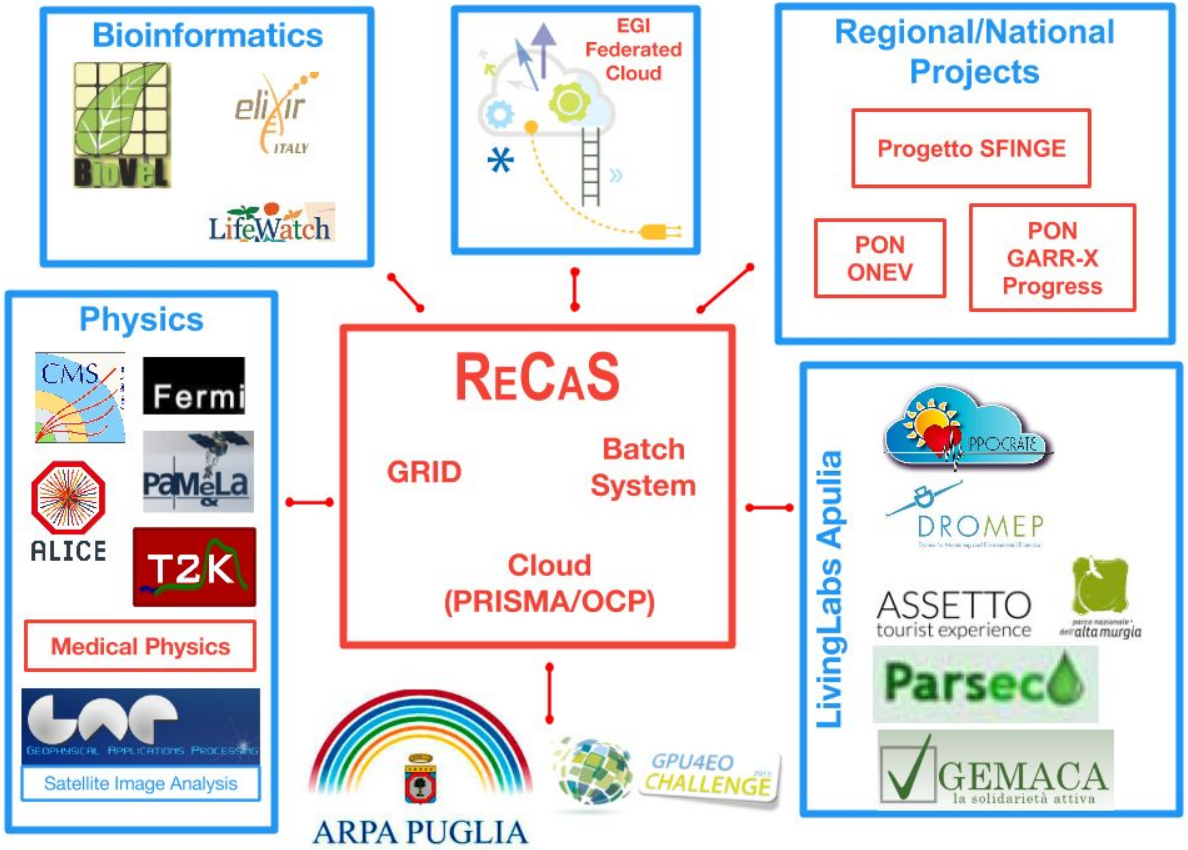
Cloud@ReCaS-Bari: setup

- 2 controller nodes
  - Expose service api endpoints
  - HAproxy Load-Balancer
  - Provide a single entry-point to the cloud services
- 34 compute nodes
  - 12 servers 32 cores Intel Xeon @2GHz, 250GB RAM
  - 22 servers 64 cores AMD Opteron, 250GB RAM
- 6 Storage Server
- RabbitMQ cluster (3 nodes)
- MySQL database (Master/Slave configuration)
- Puppet is used for the automatic installation/configuration of the services

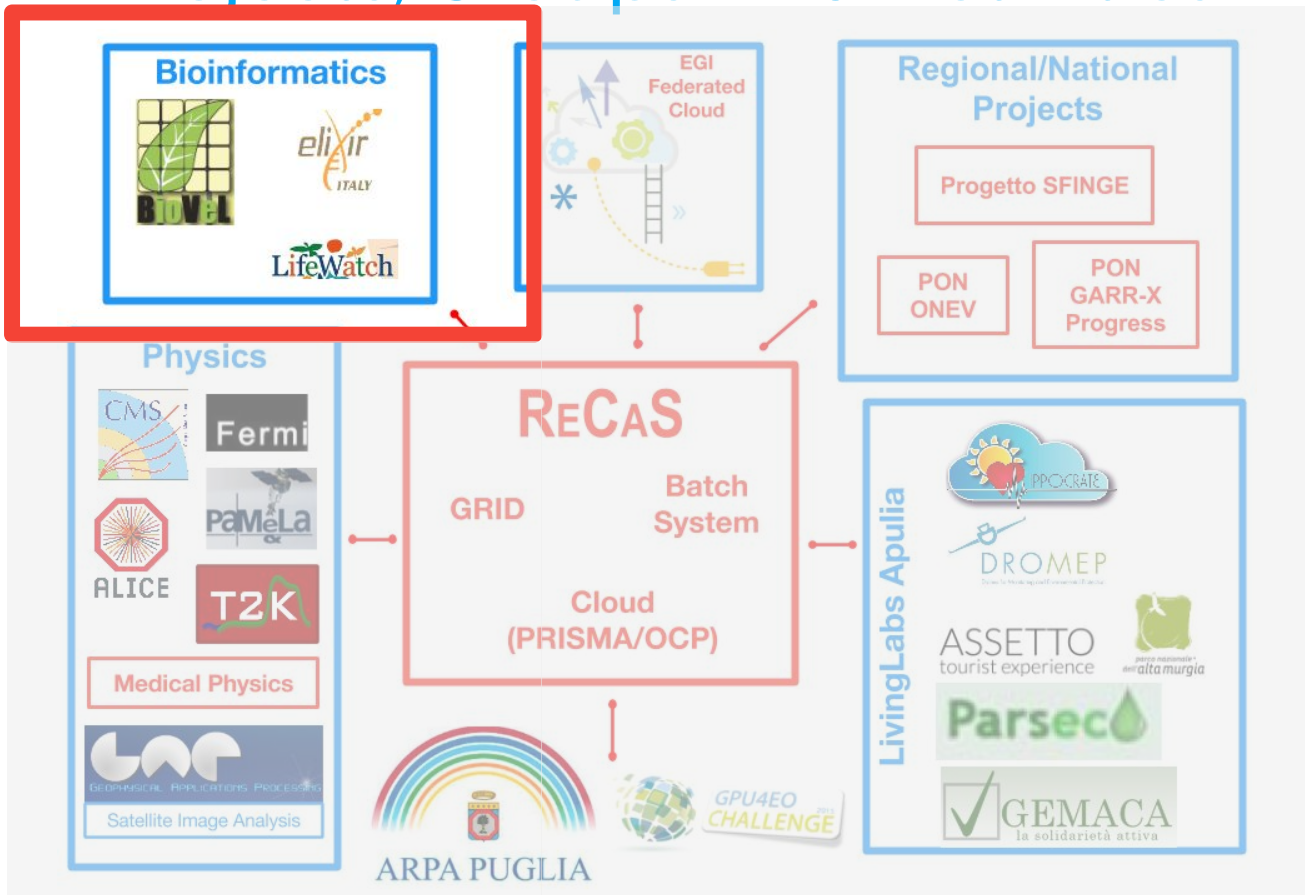
Cloud@ReCaS-Bari: numbers

- ~ 700 running Virtual Machines
- ~ 400 virtual volumes (~ 100 TB)
- 143 Active Projects
- ~ 290 users (Public Administrations, Enterprises, Research, Academia)

Projects, Groups And Activities



Projects, Groups And Activities



Why Cloud For Bioinformatics (And Training) ?

- Easier to deploy common tools (SFTP servers, analysis tools)
- Elasticity and scalability
- Easier to manage and share data
- Easier to adapt infrastructure to needs of classroom (number of users, tools, etc)
- Optimized usage of resources

Available Tools / Experience For Bioinformatics

Workflow Management Tools

LONI Pipeline
Taverna
Galaxy (web based)

Analysis Tools

MrBayes, Blast, ITK, FSL, GSNAP,
BioPython, R ,Tango, Bowtie …

Applications

BioMaS (Bioinformatic analysis of metagenomic ampliconS)
MSA-PAD (Multiple DNA Sequence Alignment framework)

BioVel portal

Evolution models
Phylogenetic Inference
Metagenomics analysis
Analysis chains developed by the project available for users

ReCaS Science Gateway

Simple access to grid/cloud resources and applications (based on Liferay)

All Governed By the Job Submission Tool (JST)

User data

WebDAV STORAGE

ownCloud

Backend

Cloud

Server

EGI

Batch Cluster

Applications

Clients

Browsers

Workflow Managers

Frontend

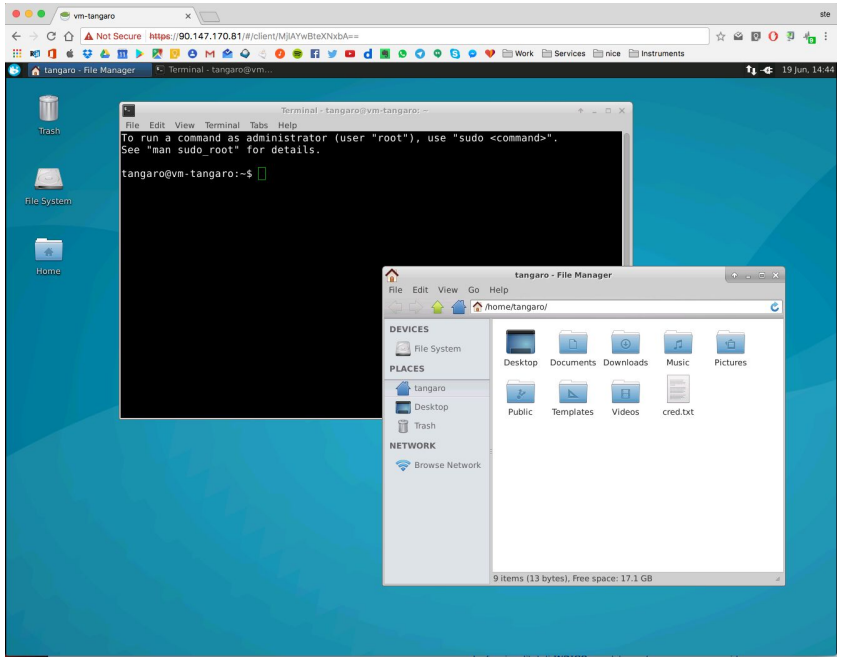
REST WS

SOAP WS

DB server

Cloud@ReCaS-Bari: advanced services

- Mesos cluster on-demand
  - Heat/Tosca template
  - Ansible roles and playbook
- Galaxy cluster on-demand
- RStudio on-demand
- Jupyterhub on-demand
- ShareLaTeX on-demand
- Dropbox-like service based on ownCloud
- Desktop as a Service (web based)



jupyter

RStudio

docker

ShareLaTeX

MESOS

ownCloud

Future/Present Perspectives - INDIGO

The European project **INDIGO DataCloud** is developing an open source data and computing platform targeted at scientific communities, deployable on multiple hardware and provisioned over hybrid, private or public, e-infrastructures:

- flexible data sharing across groups & infrastructures
- multiple sources and storage locations
- transparent network interconnections for distributed computing and storage resources
- dynamic and complex workflow management

Among the supported use-case of INDIGO there is an **on-demand one-click scalable Galaxy installation**.

INDIGO exploits Docker, Apache Mesos and OneData to manage data and application in an easy and flexible way

ONEDATA

Open source storage solution for integrating access to your data from various providers

team can easily share and process data on large scale infrastructures with the desired security level

People

M. Antonacci, D. Diacono, G. Donvito,

R. Gallitelli, R. Gervasoni, F. Giannuzzi,

A. Italiano, G. Maggi, A. Monaco,

SN, V. Spinoso, M. Tangaro,

M. Tomaiuolo, R. Valentini

Thank you
for
your attention