Commercial Clouds

Marco Landoni
marco.landoni@inaf.it

IA2 Workshop Palermo - Giugno 2022

Summary

- Introduction
- Field of application of Commercial Clouds
- Case Study 1 - The AWS call @ INAF
- Case Study 2 - The Google Cloud test @ INAF
- Conclusion

Why Commercial Clouds ?

- An increasing number of **mid/small size** problems are arising in our Institution (and around).
- Those scenarios involves the use of resources typically larger than an high-end workstation but not justified in the context of large HPC/HTC facilities (e.g. CINECA, Pleiadi, etc)
- An overall buffer would allow to **better allocate resources** w.r.t. each scenarios.

Typical examples

- Large embarrassingly parallel simulations , data reduction that **DO NOT require message passing**
- **Typically** based on code easily portable on Docker containers
- Queueable with Batch Jobs or similar.

Few state-of-the-art offered services

- Serverless Computing (e.g. Lambda Function, Cloud Function)
- Machine Learning as a Service (MLaaS)
- Accellerated computing as a Service (or in-hardware computation)
- Containerisation, execution and orchestration.

Different types of commercial clouds

Case 1 - The AWS call @ ICT INAF

- In 2020, the INAF ICT issued a call to the community for using computational power offered with AWS.
- 18 projects up to now have been approved.
- Services used: Mainly compute, jobs and GPU.

Access to the resources from the community

On demand call
Rapid access to resources (2 days typical)
Support offered

AWS Experience - a review

Among the projects approved, a number of that seems to perfectly represent the philosophy behind the use and approach of Cloud Computing (not only commercial) in general:.

- Use of Windows Based ray-tracing software (Zemax) in the context of ELT Maory project. Ultra High end cluster of workstations sporadically required on demand according to the foreseen reviews of the project. (Magrin et al 2021) **Successfully completed**
- Development and scale test environment for GPU based astrophysical codes (Covino et al 2022)
  Successfully completed
- Use of GPU as IaaS for simulations (Fabio Rossi et al 2021) **Successfully completed**
- Montecarlo Markov Chain simulation (EMCEE) in the context of gravitational waves data analysies (Salafia et al 2021, Nature) **Successfully completed**

AWS Experience - a review of projects in progress

There are also a number of projects, still ongoing, approved that are exploiting the platform.

- Machine Learning Assisted Cosmology (Granett et al 2022). Use of GPU nodes for the training of Deep Learning Neural Networks with Tensorflow (on the context of Euclid)
- Machine Learning methods for GRB detection with AGILE (Parmiggiani et al 2022)

- Analysis of Juno/JIRAM limb spectra (use of parallel Jobs with containers for data reduction, Castagnoli et al 2022)
- Large GRB simulations for the Cherenkov Telescope Array (parallel Jobs, Nava et al 2021)

Used core/hrs

- Difficult to assess in terms of vCPU/hrs due to the heterogeneity of the offered machine
- However, applying a roughly estimation about 600.000 equivalent vCPU/hrs have been used
- Few kEUR spent, more available

Main technologies used

- AWS EC2 (especially with Spot Instances) and AWS Jobs for orchestrating the runs and the Docker containers.

Storage

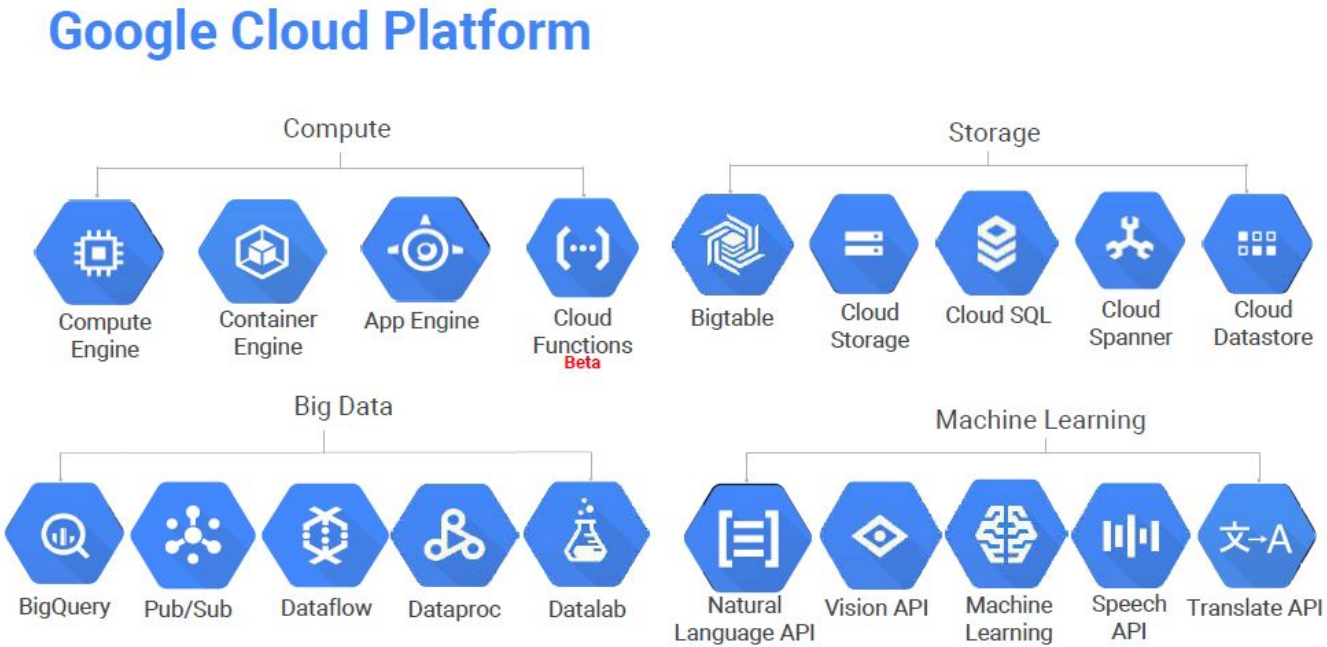
- No particular request (at the moment) for short/medium term storage
- IN general terms, Storage on the Cloud is easy and reasonable for long term, persistent storage that is not accessed (e.g. long term preservation)
- The cost for injecting the data in the sotrage (S3, EFS,...) is in general cheap (few cents for a GB) but egress are expensive (up to 0.10 EUR/GB on the cold line).
- A good solution of ultra-long term data preservation ?

Papers published

- 5 Papers (1 Nature) published in the first run (2020-2021) with the use of AWS platform, using only few kEUR of the allocated budget.
- Expected from the in-progress projects at least further 2-3 refereed papers on main journals.
- A successful experience that the ICT, according to the budget, plans to maintain in the foreseeable future.

More details: https://www.ict.inaf.it/computing/gcloud/

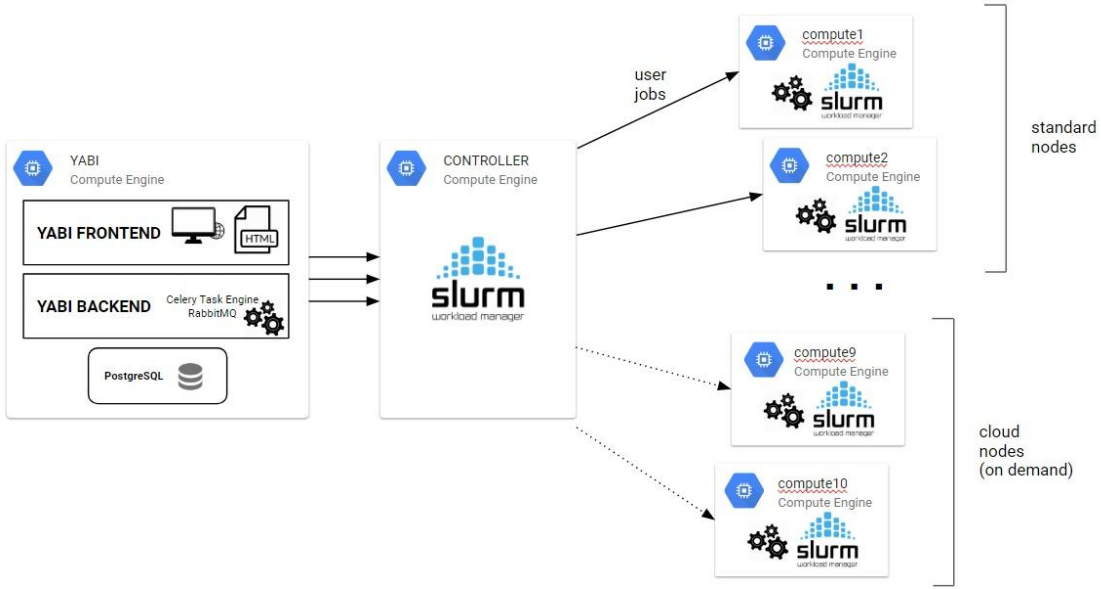
The Proof of Concepts with Google Cloud Platform



HTC workload

Compute engine has been used (like AWS EC2) for tackling HTC workload on the platform.

Yabi + GOFIO on GCP (Bignamini et al)



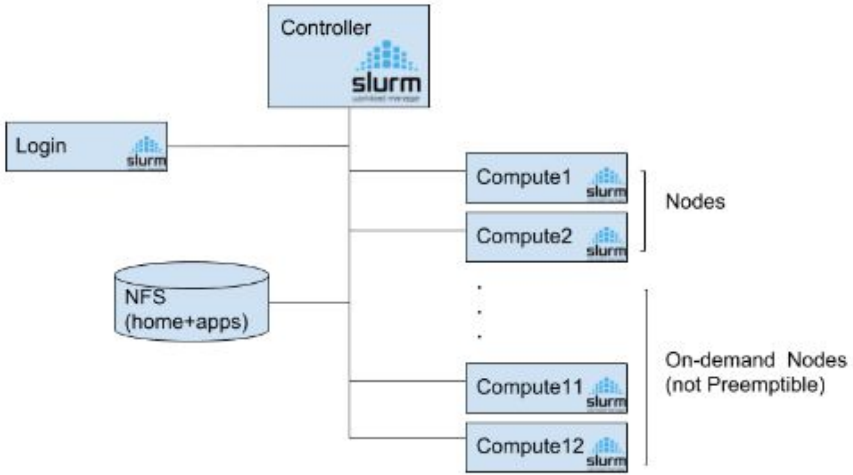
HTC + GPU Workload

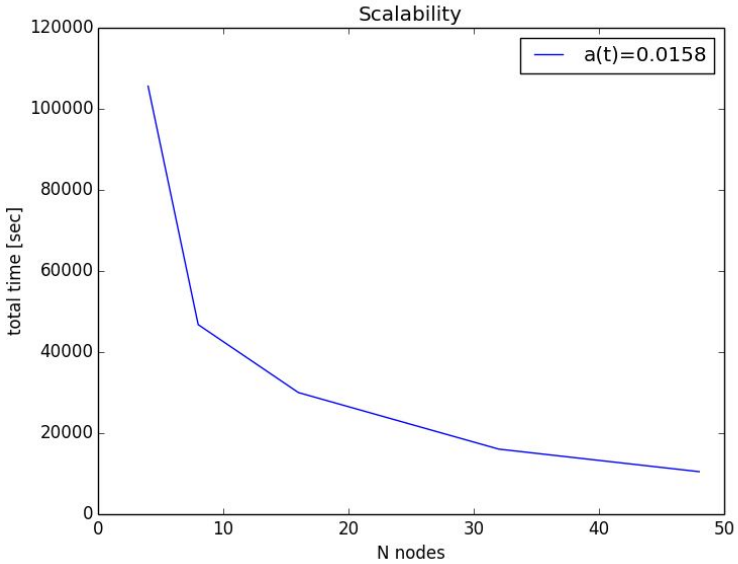
Adaptive Optics software and simulation on GCP (PASSATA, Agapito et a 2017)

- ➔ Architecture has been implemented successfully with **Managed Group + Preemptible instances**
- ➔ **SLI configuration of the GPUs** performed well with different hardware configuration (P100, Tesla K80, etc
- ➔ Overall price **low** (couple of kEUR for a whole MAORY simulation which cover about 1yr/work)
- ➔ GOOD candidate especially for GPU bounded application

HPC limitation

No HPC workload or, in general, with high message passing (Taffoni et al 2019)





EOSC & others.

- Do not forget that Cloud Computing is a computational paradigm from ideas to implementations to **papers.**
- No general preference is given for a commercial one. They are ready to use without overheads on proposals and can be accounted in the budget of a project.
- However, open and alternative, sometimes free, solution are (or will be) available in the future (GARR also is testing some Cloud Computing services)

Conclusions

- Cloud computing (not only commercial) will become more important in the near future according to the increase demand of small/medium size problems required by scientist.
- With a small budget, can be used along with state-of-the-art HPC facilities (e.g. CINECA, PLEIADI, Tecnolopo, etc.) to better allocate resources, reduce the queue and adapt computational problem to each proper environment.
- Keep an eye also on non-commercial products that are becoming more and more interesting (stay tuned)

Questions ?