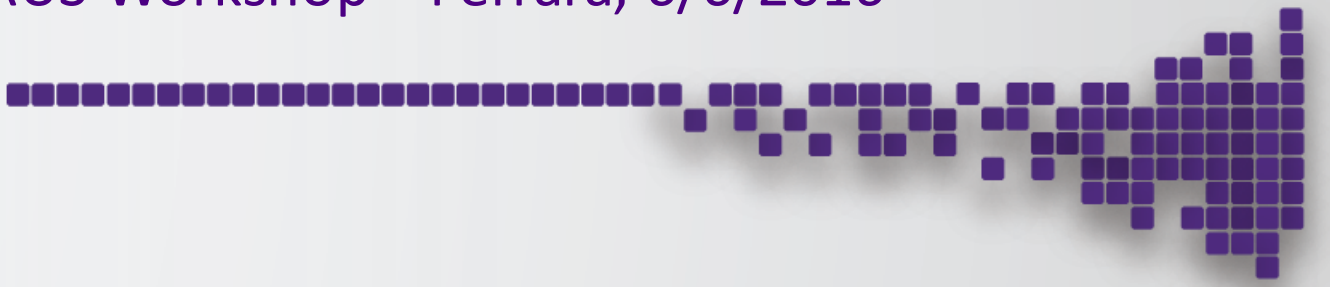
INDIGO-DataCloud:
Handling Compute and Data Intensive Scientific Applications in Cloud Infrastructures

TORUS Workshop – Ferrara, 6/6/2016



INDIGO - DataCloud

RIA-653549

Davide Salomoni, INFN-CNAF

INDIGO-DataCloud Project Coordinator

davide.salomoni@cnaf.infn.it

INDIGO-DataCloud is co-founded by the Horizon 2020Framework Programme

INDIGO-DataCloud

- **An H2020 project** approved in January 2015 in the EINFRA-1-2014 call
  - 11.1M€, 30 months (**from April 2015 to September 2017**)
- **Who**: **26 European partners** in 11 European countries
  - Coordination by the Italian National Institute for Nuclear Physics (INFN)
  - Including developers of distributed software, industrial partners, research institutes, universities, e-infrastructures
- **What**: **develop an open source Cloud platform** for computing and data ("DataCloud") tailored to science.
- **For**: **multi-disciplinary scientific communities**
  - E.g. structural biology, earth science, physics, bioinformatics, cultural heritage, astrophysics, life science, climatology
- **Where**: deployable on **hybrid (public or private) Cloud infrastructures**
  - INDIGO = **IN**tegrating **D**istributed data **I**nfrastructures for **G**lobal Expl**O**itation
- **Why**: answer to the technological **needs of scientists** seeking to easily exploit distributed Cloud/Grid compute and data resources.

INDIGO: Software for Science

INDIGO - DataCloud

INDIGO-DataCloud **develops an Open Source data and computing platform for science** provisioned over private, public or hybrid e-infrastructures.

By **filling gaps** of current Cloud technologies, INDIGO-DataCloud helps scientists, software developers, resource providers and e-infrastructures to **efficiently exploit computing, data and network technologies**:

Better Software for Better Science.

From the Paper “Advances in Cloud”

- **EC Expert Group Report on Cloud Computing**, http://cordis.europa.eu/fp7/ict/ssai/docs/future-cc-2may-finalreport-experts.pdf

To reach the full promises of CLOUD computing, major aspects have not yet been developed and realised and in some cases not even researched. Prominent among these are **open interoperation across (proprietary) CLOUD solutions at IaaS, PaaS and SaaS levels**. A second issue is **managing multitenancy** at large scale and in heterogeneous environments. A third is **dynamic and seamless elasticity** from in- house CLOUD to public CLOUDs for unusual (scale, complexity) and/or infrequent requirements. A fourth is **data management in a CLOUD environment**: bandwidth may not permit shipping data to the CLOUD environment and there are many associated legal problems concerning security and privacy. All these challenges are opportunities towards a more powerful CLOUD ecosystem.

[…] **A major opportunity for Europe involves finding a SaaS interoperable solution across multiple CLOUD platforms. Another lies in migrating legacy applications without losing the benefits of the CLOUD, i.e. exploiting the main characteristics, such as elasticity etc.**

INDIGO Addresses Cloud Gaps

- **INDIGO focuses on use cases presented by its scientific communities** to address the gaps identified by the previously mentioned EC Report, with regard to a number of areas, such as:
  - Redundancy / reliability
  - Scalability (elasticity)
  - Resource utilization
  - Multi-tenancy issues
  - Lock-in
  - Moving to the Cloud
  - Data challenges: streaming, multimedia, big data
  - Performance

The main gaps, in summary:

1. **Open interoperation across (proprietary) Cloud solutions at IaaS, PaaS and SaaS levels.**
2. **Managing multitenancy at large scale and in heterogeneous environments.**
3. **Dynamic and seamless elasticity from in-house Cloud to public Clouds for unusual (scale, complexity) and/or infrequent requirements.**
4. **Handling data management in a Cloud environment.**

INDIGO for Scientists

- INDIGO's user-oriented access services, efficient exploitation of available resources and integrated access to distributed data allow scientific communities to

Access data and use resources as a "big pool" of computing and storage, without the need to know their type or location.

- This happens directly through INDIGO services and does not require writing specialized software.

INDIGO for e-Infrastructures and Service Providers

- INDIGO's user-oriented access services, efficient exploitation of available resources and improved functionalities in open source Cloud software allow resource providers, resource centers and Cloud infrastructures to

Provide tools and optimal exploitation of distributed resources and to offer new, advanced services on top of the INDIGO components.

- The INDIGO software is an extensible enabling technology for research infrastructures, SP, ESFRI projects and similar initiatives.

INDIGO and other European Projects

- The INDIGO services are being developed according to the requirements collected within many multidisciplinary scientific communities, such as **ELIXIR, WeNMR, INSTRUCT, EGI-FedCloud, DARIAH, INAF-LBT, CMCC-ENES, INAF-CTA, LifeWatch-Algae-Bloom, EMSO-MOIST**. However, they are implemented so that they can be easily reused by other user communities.
- INDIGO has strong relationships with complementary initiatives, such as **EGI-Engage** on the operational side and **AARC** with respect to AuthN/AuthZ policies. Users of EC-funded initiatives such as **PRACE** and **EUDAT** are also expected to benefit from the deployment of INDIGO components in such infrastructures.
- Several **National/Regional infrastructures** are covered by the 26 INDIGO partners, located in 11 European countries.
- INDIGO is mentioned in a recent **Important Project of Common European Interest (IPCEI)** proposal for the exploitation of HPC and HTC resources at national, regional and European levels.

Key Dates

- INDIGO-DataCloud started on April 1st, 2015.
- The first internal (beta) release was ready by March 2016, with first demos shown at: EGI Conference Nov 2015 (Bari), CloudScape 8-9 March 2016 (Brussels), and 4-5/4/2016 at the "INDIGO Champions meet Developers" meeting (Amsterdam).
- 4-6 May 2016, Frascati: All-Hands, Collaboration Board and Technical Board meetings.
- **First public release: 1st August, 2016**
- **Mid-term review by the EC** scheduled on 19-20 September 2016 in Bologna.
- Second public release due by March 2017.
- The project will end on September 30th, 2017.

- See the path for our releases at https://www.indigo-datacloud.eu/indigo-roadmap.

Latest activities

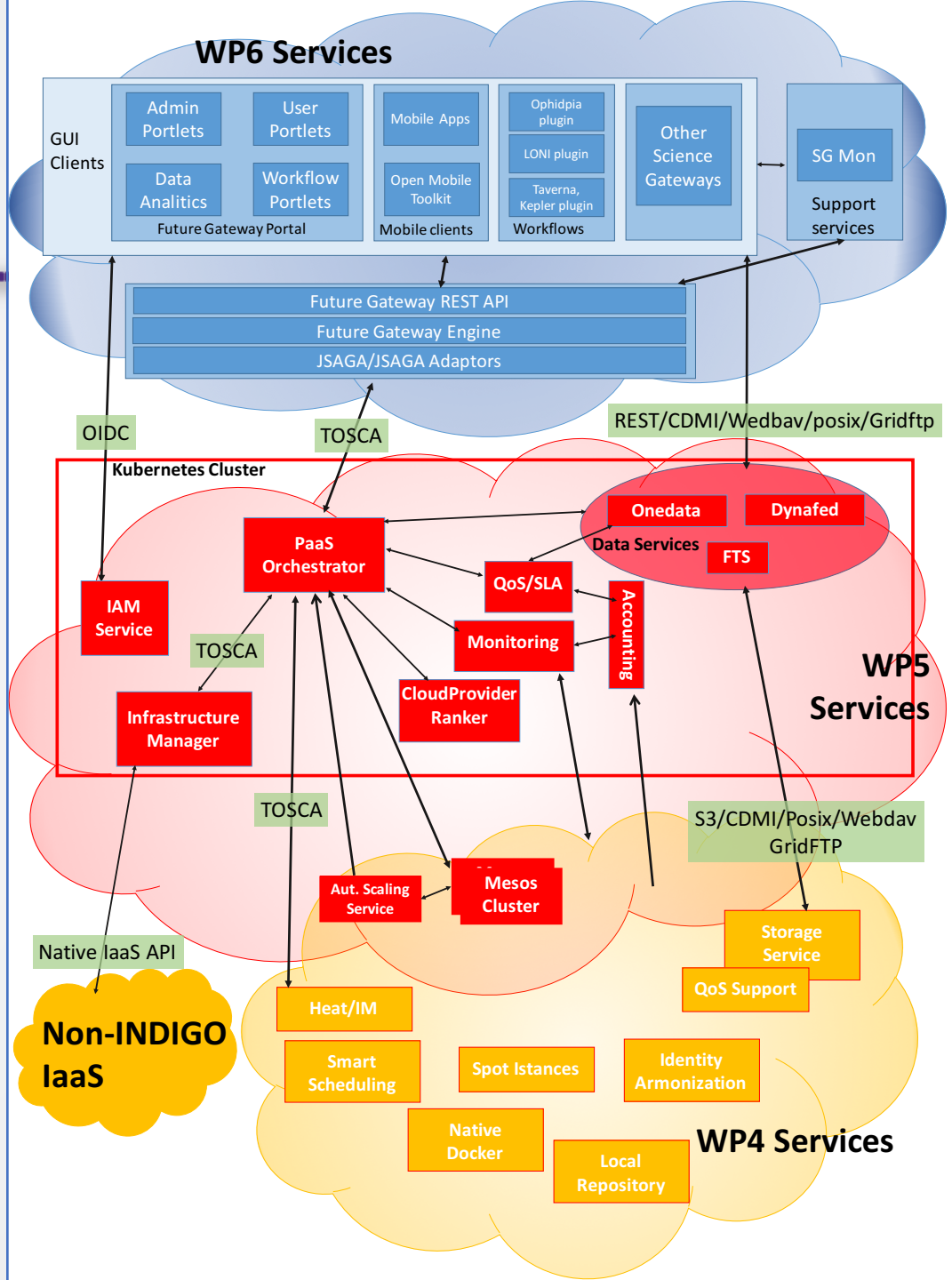
INDIGO - DataCloud

- Full definition of the INDIGO Architecture recently published on arXiv (http://arxiv.org/abs/1603.09536).
- More than 120 ”user stories” collected.
- Submission of articles and presentations to many conferences (as an example, we recently submitted 11 abstracts to CHEP2016 only).
- 10+ concrete implementations being worked on right now by our scientific community champions, together with developers and high-level experts.
  - For instance: use cases by EuroBioImaging, LifeWatch, Large Binocular Telescope, Cherenkov Telescope Array, Elixir, Climate modeling, High-Energy Physics, EMSO, WeNMR/INSTRUCT.
  - With a number of publications in the pipeline.
- Several external collaborations and contributions to upstream projects (e.g. OpenStack and OpenNebula).

The INDIGO Architecture

The **INDIGO Architecture** and all its components fully support the same **FAIR** principles proposed for scientific data management:

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable



The INDIGO solutions

- **The INDIGO architecture can be seen as providing:**
  - **Site-level solutions**
  - **Data solutions**
  - **Automated solutions**
  - **User-level solutions**
- All of them integrate in a consistent global framework. Frequently a given solution spans multiple INDIGO WPs.
- There are many details “behind the scenes”. They are essential and addressed by our Work Packages. But let’s focus here on a bird’s eye view from a practical perspective.

Site-level solutions

INDIGO - DataCloud

- New scheduling algorithms for open source Cloud frameworks.
  - Both **fair-share scheduling** and **spot instances**
- Full support for containers, *with or without Docker*.
- Dynamic partitioning of batch vs. Cloud resources.
- Storage QoS and data lifecycle support.
- Support for external infrastructures.
- Automated synchronization of dockerhub repos with the local repository of open source Cloud frameworks.
- Improved automation at IaaS level, based on TOSCA

Data solutions

INDIGO - DataCloud

- Integrated local and remote Posix access for all types of resources (bare metal, virtual machines, containers).
- Transparent mapping of object storage (e.g. Ceph, S3) to Posix.
- Transparent data caching and replicas.
- Transparent gateway to existing filesystems (e.g. GPFS, Lustre).
- Webdav, GridFTP, CDMI, web, fuse access.
- Dropbox-like functionalities, based on ownCloud (target: September 2016).
- Linux, Mac OS, Windows desktop support.

Automated solutions

They are typically based on TOSCA templates used to specify resource requirements, dependencies, and configuration of the services/applications (sample templates for common use cases are provided by the project).

- Selection of resources across multiple Cloud providers (e.g. depending on data location or resource requirements).
- Support for application requirements in Cloud resource allocations (e.g. for what regards InfiniBand or GPUs).
- Dynamic instantiation, automated monitoring and elasticity of long-running services.
- Dynamic instantiation, automated monitoring and elasticity of batch systems, front-ends included.
- Support for custom frameworks for porting arbitrary applications to the Cloud, with automated monitoring and scalability.
- Support for big data analysis applications (Ophidia, Spark).
- Mesos clusters transparently spanning multiple data centers (not in the first release).

User-level solutions

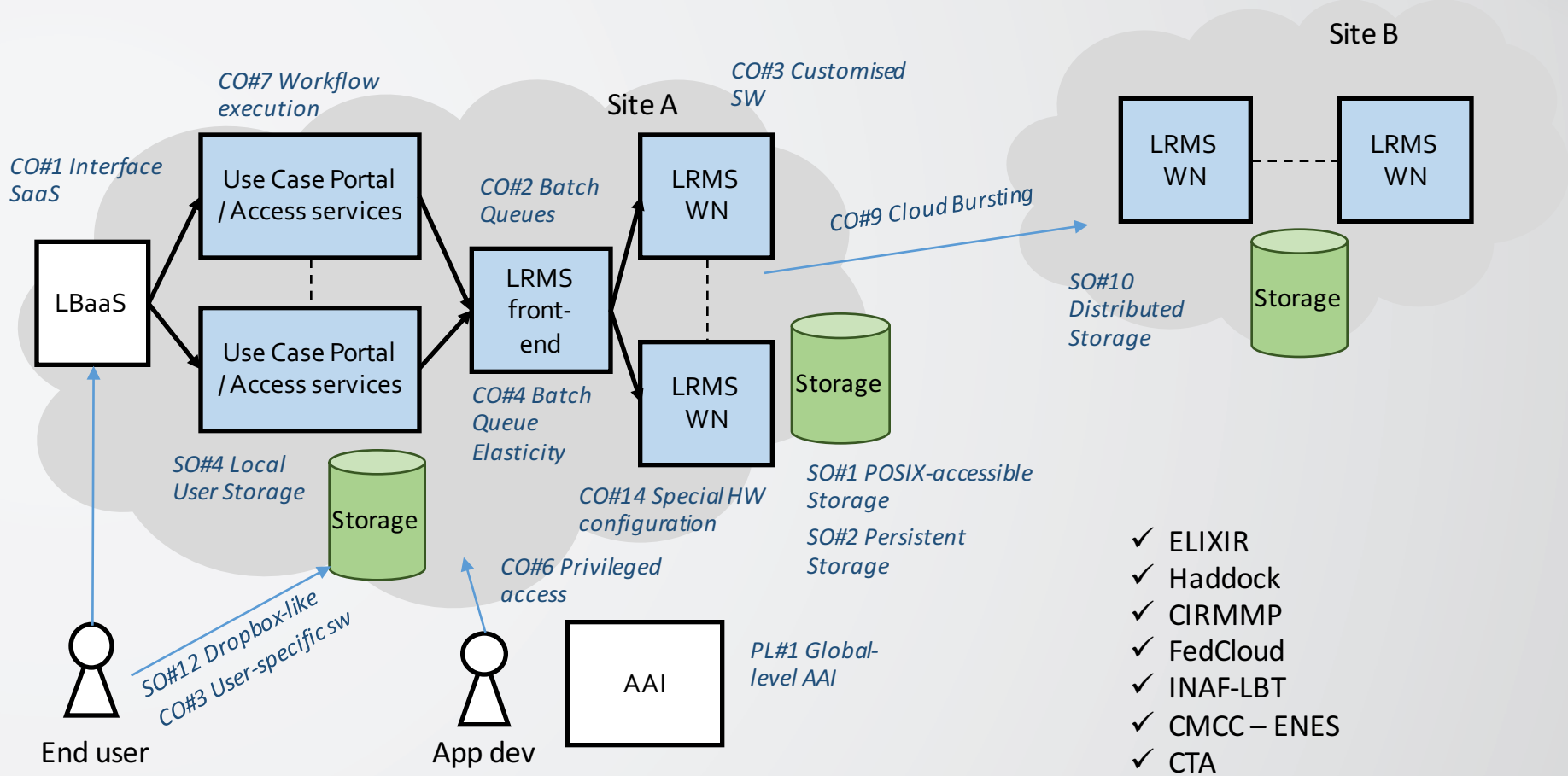
- Customizable / programmable portal (gateway) engine integrated with the features mentioned in the previous slides.
- Sample portals delivered for selected applications ("all-in-one", “plug and play” bundles).
- Mobile toolkit to access INDIGO features on mobile devices.
- AAI architecture integrated at all levels supporting X.509, SAML, OpenID Connect.

Example: Scenario #1

INDIGO - DataCloud

- **“Computational Portal as a Service”**
- A scientific community has an application (or a set of them) that should be accessed through a portal. The application:
  - Requires a dynamically instantiated batch queue as its back-end;
  - Exhibits an unpredictable workload;
  - Supports multiple access profiles;
  - Should be deployable through Cloud providers, with features such as redundancy and elasticity;
  - May require cloud bursting to other infrastructures;
  - Should support both access to external reference data and to data local to the application, which must be accessible in a distributed way.

Computational Portal as a Service



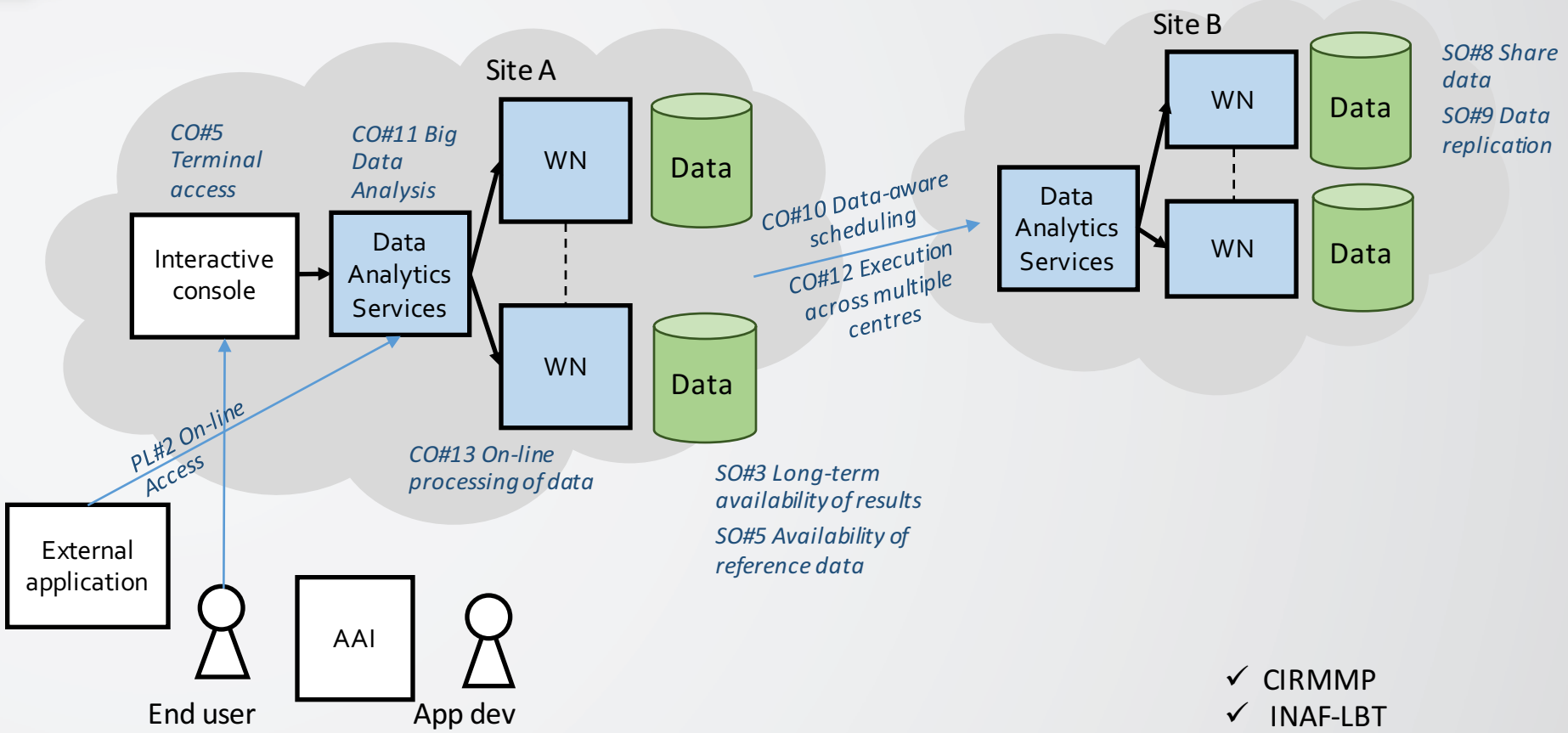
- ✓ ELIXIR
- ✓ Haddock
- ✓ CIRMMP
- ✓ FedCloud
- ✓ INAF-LBT
- ✓ CMCC – ENES
- ✓ CTA
- ✓ ALGAE – BLOSSOM

Example: Scenario #2

INDIGO - DataCloud

- **“Data Analysis Service”**
- A scientific community has a coordinated set of data repositories and software services they want to access, process and inspect. Data processing should be interactive, requiring access to a console deployed on the site where data is located. The application:
  - Consists of a console or of a scientific gateway;
  - Interacts with data and can expose programmatic services;
  - Should be deployable through Cloud providers, with features such as redundancy and elasticity.

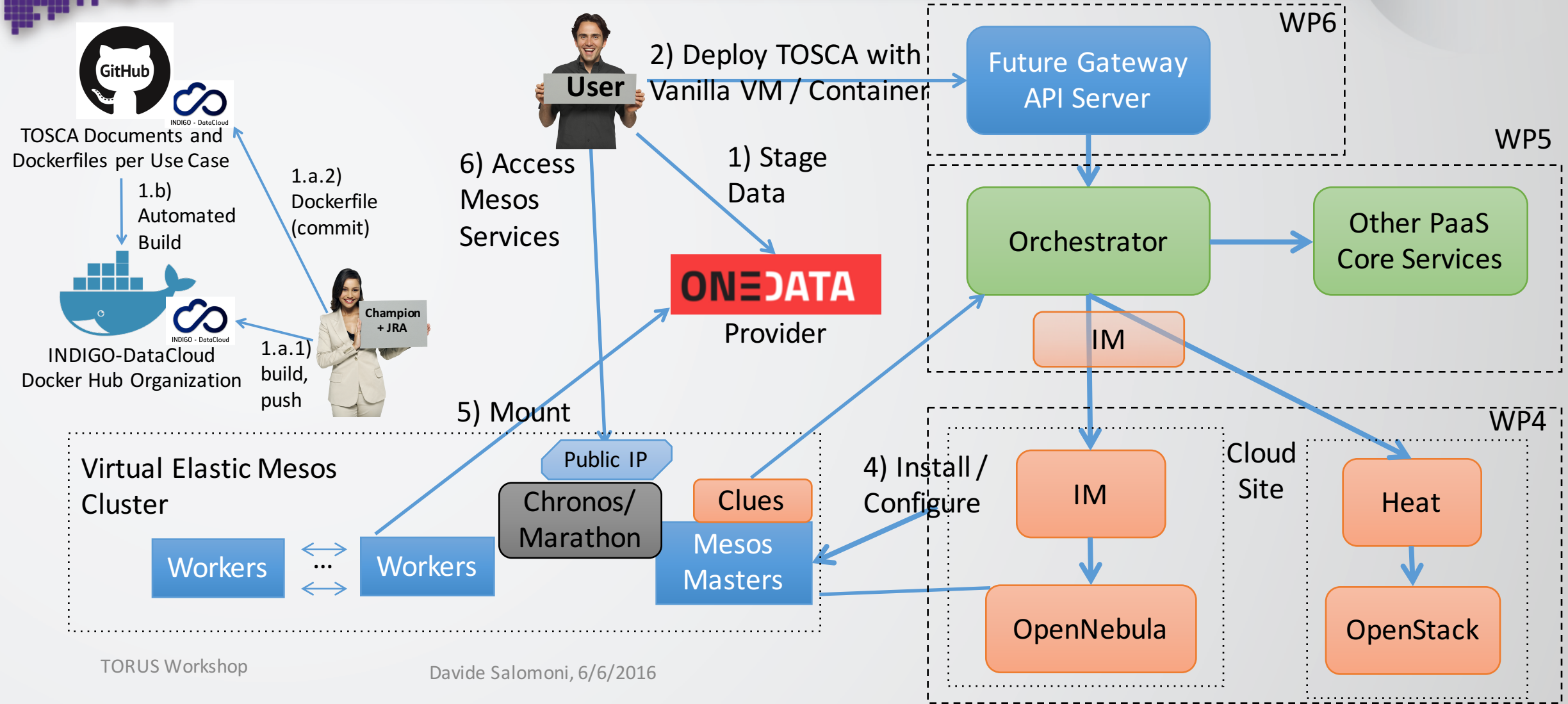
Data Analysis Service



(Other requirements from previous case also apply)

- ✓ CIRMMP
- ✓ INAF-LBT
- ✓ CMCC – ENES
- ✓ ALGAE – BLOSSOM
- ✓ INGV - MOIST

Putting several pieces together…



Exploitation (1)

- There are currently *many* activities going on with communication and collaboration with external entities, such as:
  - RDA (Research Data Alliance), specifically with the QoS and Data Lifecycle WG and Database WG.
  - SNIA (Storage Networking Industry Association).
  - National Research Networks such as GARR.
  - EC Projects such as Thor, PhenoMeNal, Beacon.
  - IBM (on TOSCA), Yahoo! (on spot instances).
  - OpenNebula (containers).
  - OpenStack (at several levels, including the newly formed OpenStack Scientific WG).
  - Yandex.
  - ESA.
  - Dissemination at the EC level (e.g. in preparation for the ICRI2016 Conference, October 2016)
  - Collaborations with non-EC institutions such as Nectar in Australia, Ohio State University and Lawrence Livermore National Lab in the US.
  - EGI Communities not directly participating to INDIGO.

Exploitation (2)

INDIGO - DataCloud

- At the INDIGO AHM (4-6/5/2016): a meeting dedicated to describing strategy and exploitation plans of the INDIGO software by the 4 INDIGO industrial partners.
- Followed by a meeting dedicated to possible exploitation of INDIGO solutions in the ESA/Copernicus context (both with ESA and with external industrial partners).
- Contacts with several other projects and organizations, such as GARR, Phenomenal (an H2020 project on data processing and analysis pipelines for molecular phenotype data generated by metabolomics applications), THOR (an H2020 project on accessing open research data), and others.
- We are going to organize dedicated training / dissemination sessions or workshops in the next future, or participate to existing ones (such as the eResearch Summer Hackfest, http://www.sci-gaia.eu/summer-hackfest/)
  - Videos are planned, and some of them are already available, see for example: https://www.youtube.com/watch?v=sEDBZFZjrvE, https://www.youtube.com/watch?v=UtbFAhvQZ40).

Conclusions

- INDIGO-DataCloud develops an **Open Source data and computing platform** provisioned over private, public or hybrid e-infrastructures that **addresses typical technology gaps** preventing easy, efficient and cost-effective exploitation of Cloud resources by scientific communities.
- **The first INDIGO release is due by end of July 2016**; a second release is foreseen by March 2017.
- We believe that **developing software for standard, easy to use, community-backed solutions is essential** to exploit the currently fragmented universe of European cloud resources and infrastructures **toward a European Open Science Cloud**.
- **Our focus** is now on delivering the components for the first release. More components and supported use cases are in the works.

https://www.indigo-datacloud.eu

Better Software for Better Science.