INDIGO - DataCloud

RIA-653549

INDIGO-DataCloud Project Status

CCR Workshop, 17/3/2016

Davide Salomoni, INFN-CNAF
INDIGO-DataCloud Project Coordinator
davide.salomoni@cnaf.infn.it

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

First things first…

TASK #1: FORZA CRISTINA, TI ASPETTIAMO IL PRIMA POSSIBILE!

Doina Cristina Aiftimiei
OpenStack for Research and Research for OpenStack
vBrownBag TechTalks at OpenStack Summit Tokyo 2015 - S

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.

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.

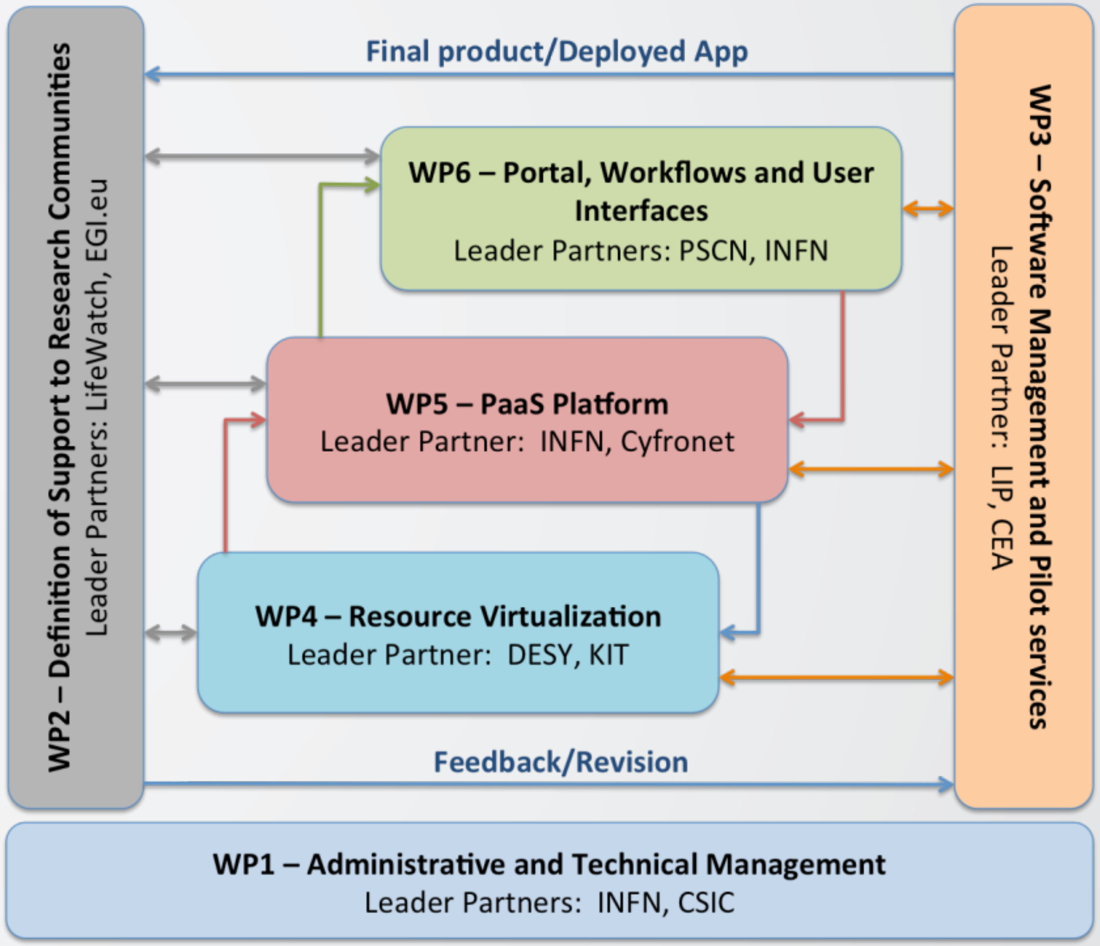
[…] **A major opportunity for Europe involves finding a SaaS interoperable solution across multiple CLOUD platforms. Another lies in migrating legacy applications without losing the benefits of the CLOUD, i.e. exploiting the main characteristics, such as elasticity etc.**

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:
  - Redundancy / reliability
  - Scalability (elasticity)
  - Resource utilization
  - Multi-tenancy issues
  - Lock-in
  - Moving to the Cloud
  - Data challenges: streaming, multimedia, big data
  - Performance
- **Reusing existing open source components** wherever possible and **contributing to upstream projects** (such as OpenStack, OpenNebula, Galaxy, etc.) for sustainability.

Work Packages

INDIGO - DataCloud



WP2 – Definition of Support to Research Communities
Leader Partners: LifeWatch, EGI.eu

Final product/Deployed App

WP6 – Portal, Workflows and User Interfaces
Leader Partners: PSCN, INFN

WP5 – PaaS Platform
Leader Partner: INFN, Cyfronet

WP4 – Resource Virtualization
Leader Partner: DESY, KIT

Feedback/Revision

WP3 – Software Management and Pilot services
Leader Partner: LIP, CEA

WP1 – Administrative and Technical Management
Leader Partners: INFN, CSIC

Key Dates

- INDIGO-DataCloud started on April 1st, 2015
- The first internal beta release is due by March 2016
- 4-5 April 2016, Amsterdam: meeting between the "WP2 Champions" (see later) and the Development Work Packages to define technical implementation of the use cases following the first beta release.
- 4-6 May 2016, Frascati: All-Hands, Collaboration Board and Technical Board meetings.
- **First public release due by July, 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

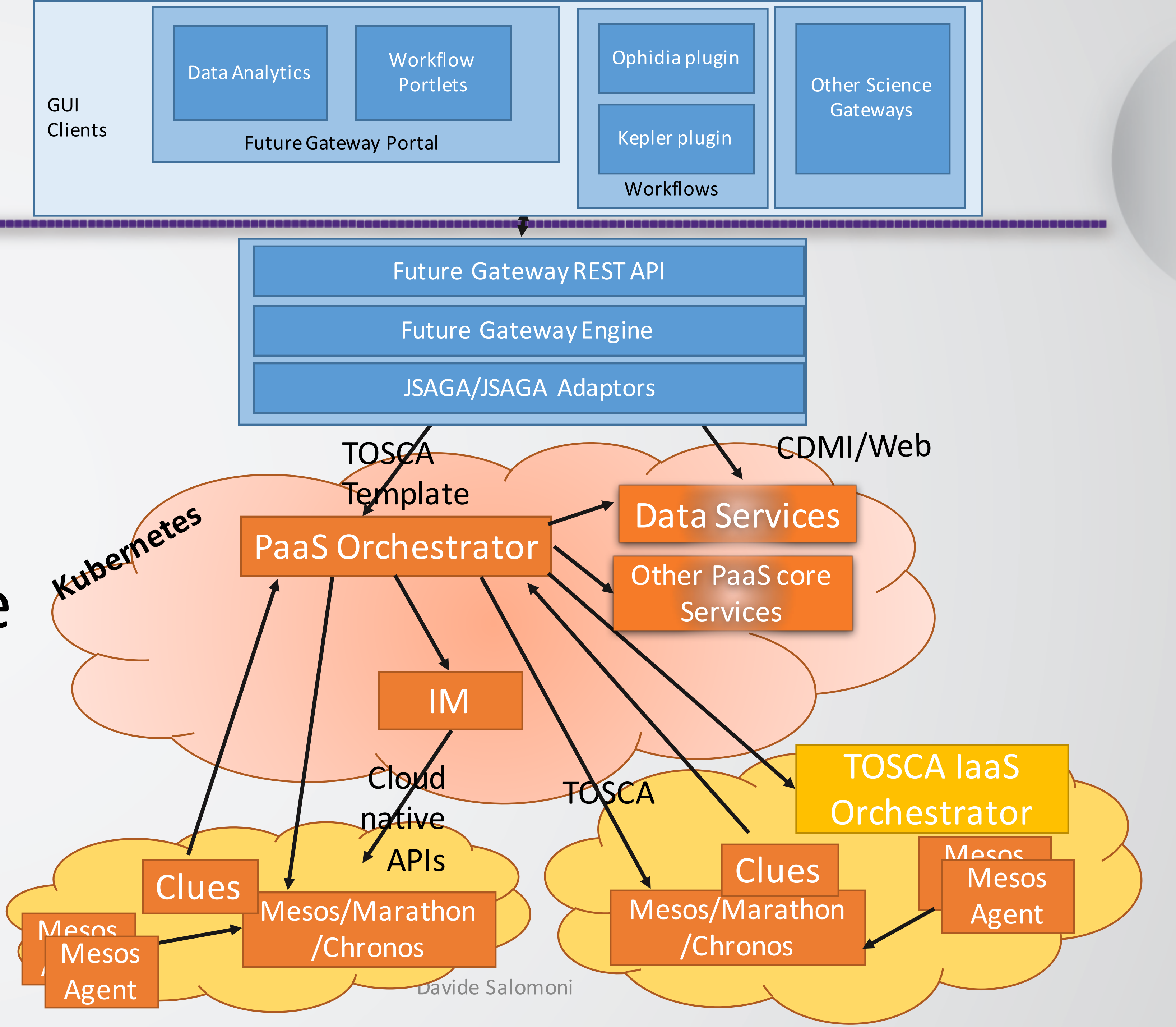
User Communities

- **Requirements from user communities** belonging to the INDIGO Consortium have been analyzed in Deliverable 2.1 (“Initial Requirements from Research Communities”), see https://owncloud.indigo-datacloud.eu/index.php/s/86iqWJhSC9dxwai
- Some more detailed **specification of the use cases** can be found in Deliverable D2.3 (“Specification of Use Cases for Testing and Validation”), see https://owncloud.indigo-datacloud.eu/index.php/s/2MvtRY36XgDChV2
- An **updated version of the requirements** can be found in Deliverable D2.4 (“Confirmation of Support to Initial requirements and extended list”), https://owncloud.indigo-datacloud.eu/index.php/s/pLCB3uSwlmgtjHP.
- Research Communities covering a **wide and significant spectrum of areas and expertise** are represented through the participation of relevant institutions associated to ESFRIs or EIROs:
  - Biological and Medical Sciences: EuroBioImaging –BBMRI (UPV), ELIXIR (CNR), INSTRUCT (U.Utrecht, CIRMMP)
  - Social Sciences and Humanities: DARIAH (RBI), DCH-RP (ICCU)
  - Environmental and Earth Sciences: LifeWatch (CSIC), EMSO (INGV), ENES (CMCC)
  - Physical Sciences: LBT, CTA (INAF) [+conduit to WLCG+HEP from CERN]
- **Each research community has nominated a “Champion”** who will follow in detail the adoption of INDIGO solutions for his own use cases.

Key achievements so far

- Definition and adoption of support tools for project management and tracking (based on OpenProject), mailing list (sympa), web (Joomla), dropbox-like (owncloud), agenda (indico).
- Definition of a Continuous Integration infrastructure (based on Jenkins). Definition of several development testbeds.
- 15 Deliverables submitted to the EC (13 of them public and available on the website at https://www.indigo-datacloud.eu/documents-deliverables), documenting all phases of the project.
- Participation to several International workshops and conferences (among them: European Open Science Cloud initiatives, RDA, EGI, ISC, SuperComputing, AAAS, OpenNebula Conference, OpenStack Summit, ISGC)
- Social media presence (Twitter, Facebook)
- 40 repos of INDIGO software and dockerfiles are currently present on github.
- Awareness of the INDIGO solutions is growing, judging by the numerous requests for information and collaboration we are getting from SPs, e-infrastructures, industries and research communities *outside* the INDIGO Consortium.
- We have active collaborations with key industrial players such as IBM and Yahoo!
- SNIA recently accepted our proposals/commits about CDMI extensions.
- We started up a new interest group within RDA to discuss storage QoS and data life cycle management.

INDIGO-DataCloud General Architecture

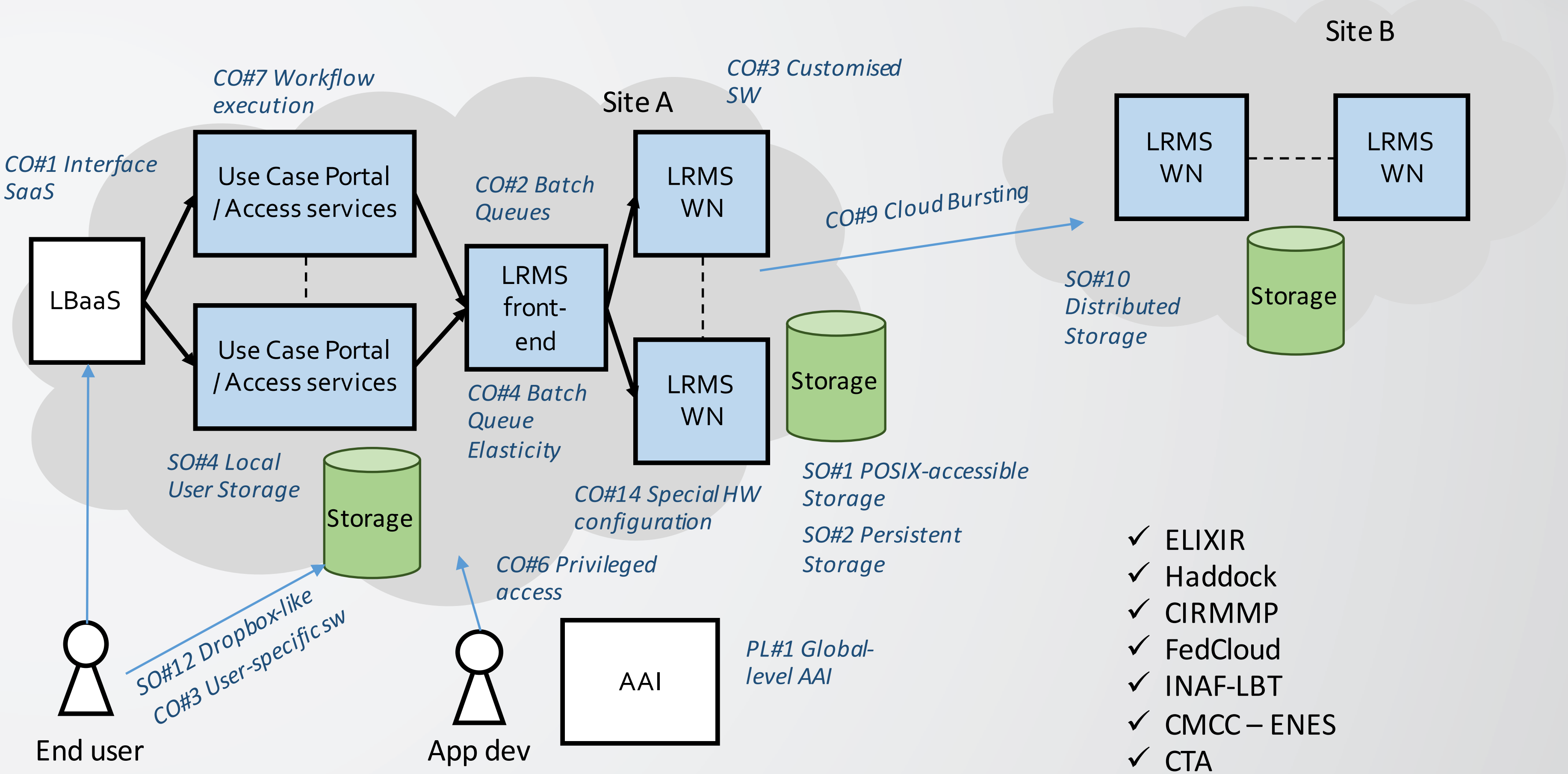


INDIGO - DataCloud

Example: Scenario #1

- **“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

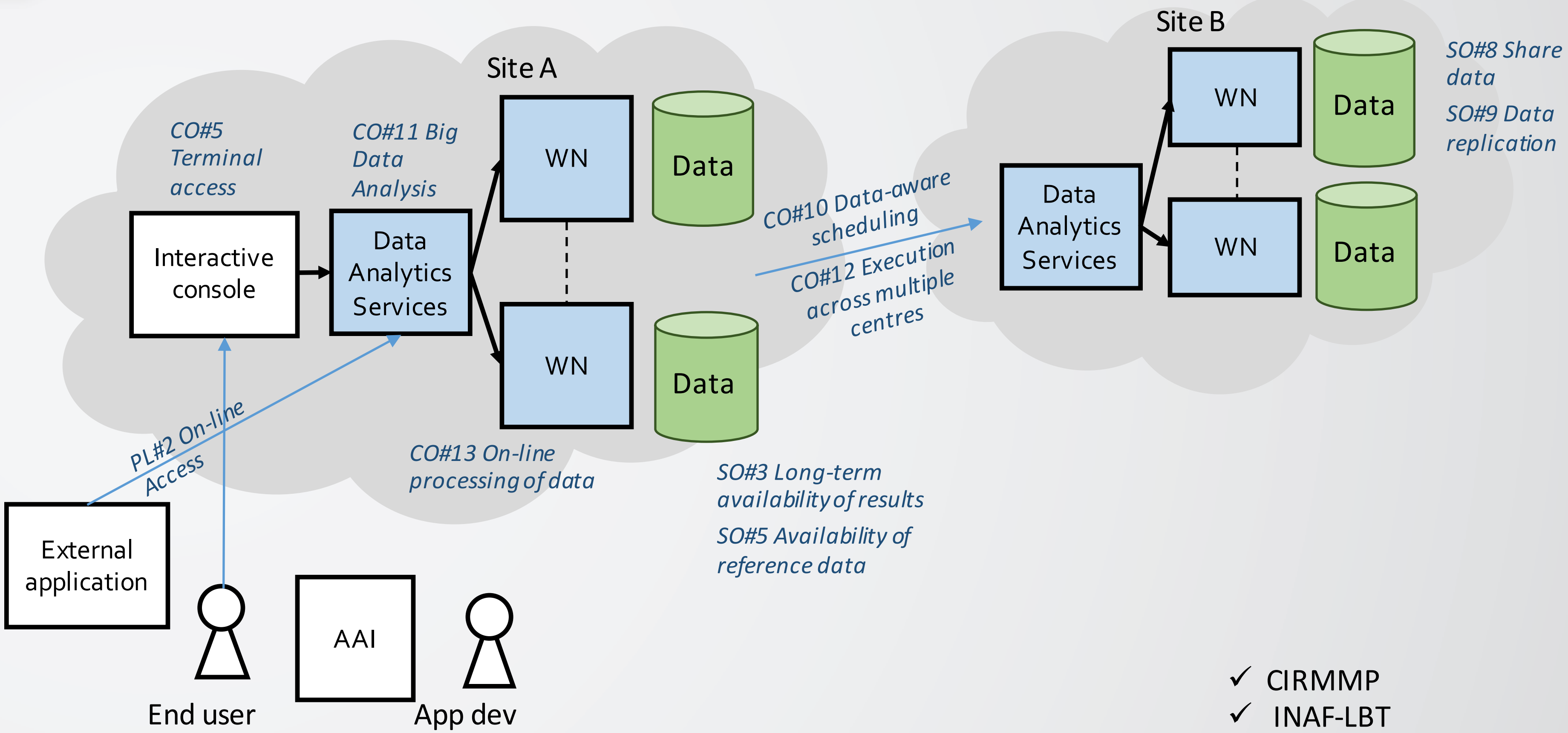


- ✓ ELIXIR
- ✓ Haddock
- ✓ CIRMMP
- ✓ FedCloud
- ✓ INAF-LBT
- ✓ CMCC – ENES
- ✓ CTA
- ✓ ALGAE – BLOSSOM

Example: Scenario #2

- **“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.

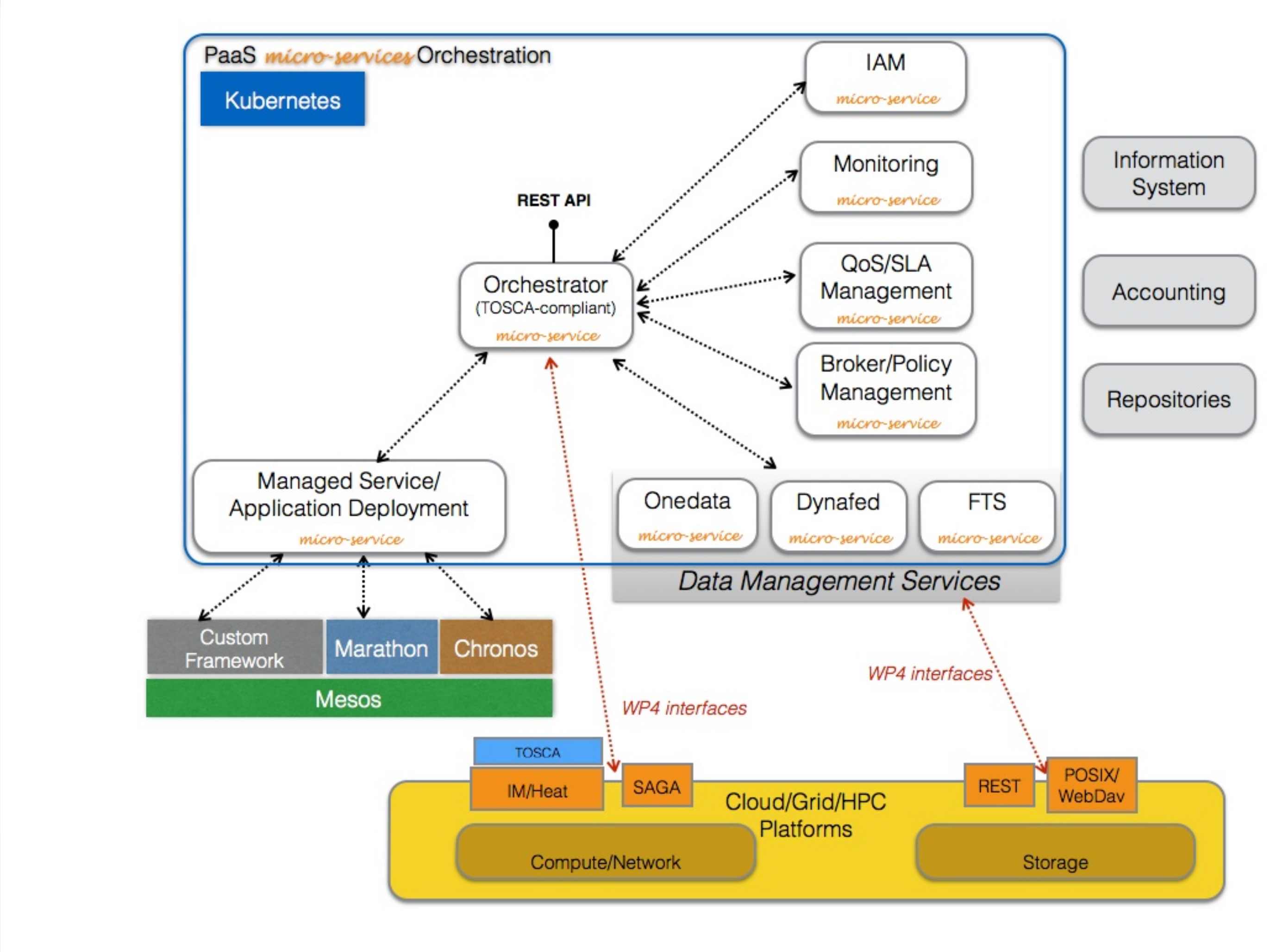
Data Analysis Service



(Other requirements from previous case also apply)

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

Key INDIGO PaaS Components



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 the recent **Important Project of Common European Interest (IPCEI)** for the exploitation of HPC and HTC resources at national, regional and European levels.

Thank you

https://www.indigo-datacloud.eu

Better Software for Better Science.

Back-up Slides with key INDIGO Iaas, PaaS and SaaS features

IaaS Features (1)

- **Improved scheduling for allocation of resources** by popular open source Cloud platforms, i.e. OpenStack and OpenNebula.
  - Enhancements will address both better scheduling algorithms and support for spot-instances. The latter are in particular needed to support allocation mechanisms similar to those available on public clouds such as Amazon and Google.
  - We will also support dynamic partitioning of resources among “traditional batch systems” and Cloud infrastructures (for some LRMS).
- **Support for standards in IaaS resource orchestration engines** through the use of the TOSCA standard.
  - This overcomes the portability and usability problem that ways of orchestrating resources in Cloud computing frameworks widely differ among each other.
- **Improved IaaS orchestration capabilities** for popular open source Cloud platforms, i.e. OpenStack and OpenNebula.
  - Enhancements will include the development of custom TOSCA templates to facilitate resource orchestration for end users, increased scalability of deployed resources and support of orchestration capabilities for OpenNebula.

IaaS Features (2)

- **Improved QoS capabilities of storage resources**.
  - Better support of high-level storage requirements such as flexible allocation of disk or tape storage space and support for data life cycle. This is an enhancement also with respect to what is currently available in public clouds, such as Amazon Glacier and Google Cloud Storage.
- **Improved capabilities for networking support**.
  - Enhancements will include flexible networking support in OpenNebula and handling of network configurations through developments of the OCCI standard for both OpenNebula and OpenStack.
- **Improved and transparent support for Docker containers**.
  - Introduction of native container support in OpenNebula, development of standard interfaces using the OCCI protocol to drive container support in both OpenNebula and OpenStack.

PaaS Features (1)

- **Improved capabilities in the geographical exploitation of Cloud resources**.
  - End users need not to know where resources are located, because the INDIGO PaaS layer is hiding the complexity of both scheduling and brokering.
- **Standard interface to access PaaS services**.
  - Currently, each PaaS solution available on the market is using a different set of APIs, languages, etc. INDIGO will use the TOSCA standard to hide these differences.
- **Support for data requirements in Cloud resource allocations**.
  - Resources can be allocated where data is stored.
- **Integrated use of resources coming from both public and private Cloud infrastructures**.
  - The INDIGO resource orchestrator is capable of addressing both types of Cloud infrastructures through TOSCA templates handled at either the PaaS or IaaS level.

PaaS Features (2)

- **Distributed data federations** supporting legacy applications as well as high level capabilities for distributed QoS and Data Lifecycle Management.
  - This includes for example remote Posix access to data.
- **Integrated IaaS and PaaS support in resource allocations**.
  - For example, storage provided at the IaaS layer is automatically made available to higher-level allocation resources performed at the PaaS layer.
- **Transparent client-side import/export of distributed Cloud data**.
  - This supports dropbox-like mechanisms for importing and exporting data from/to the Cloud. That data can then be easily ingested by Cloud applications through the INDIGO unified data tools.
- **Support for distributed data caching mechanisms and integration with existing storage infrastructures**.
  - INDIGO storage solutions are capable of providing efficient access to data and of transparently connecting to Posix filesystems already available in data centers.

PaaS Features (3)

- **Deployment, monitoring and automatic scalability of existing applications**.
  - For example, existing applications such as web front-ends or R-Studio servers can be automatically and dynamically deployed in highly-available and scalable configurations.
- **Integrated support for high-performance Big Data analytics**.
  - This includes custom frameworks such as Ophidia (providing a high performance workflow execution environment for Big Data Analytics on large volumes of scientific data) as well as general purpose engines for large-scale data processing such as Spark, all integrated to make use of the INDIGO PaaS features.
- **Support for dynamic and elastic clusters of resources**.
  - Resources and applications can be clustered through the INDIGO APIs. This includes for example batch systems on-demand (such as HTCondor or Torque) and extensible application platforms (such as Apache Mesos) capable of supporting both application execution and instantiation of long-running services.

SaaS Features (1)

- **Development of an extensible mobile and web platform *API* for the integration of applications exploiting cloud resources**.
  - Through the INDIGO-developed APIs it is possible to develop web or mobile applications inheriting all the INDIGO PaaS- and IaaS-level enhancements. Some example scientific portals, tailored for selected applications, will also be provided.
- **Seamless interface for the exploitation of *Grid, Cloud and local* resources supporting also complex workflow management systems**.
  - The developed APIs make it possible for existing workflow management systems such as Kepler and Taverna to efficiently exploit *different kinds of distributed computing and storage resources*.