INDIGO-DataCloud

Davide Salomoni, INFN-CNAF
(davide.salomoni@cnaf.infn.it)

15/12/2014

Stato di INDIGO

- **INDIGO-DataCloud** (**IN**tegrating **D**istributed data **I**nfrastructures for **G**lobal expl**O**itation) è una proposta sottomessa alla call Horizon 2020 E-INFRA1 (scadenza 2/9/2014). L'INFN coordina la proposta.
- **Lo stato della proposta**: sappiamo che le sottomissioni a EINFRA-1 sono state valutate e che esiste un ranking, che tuttavia ancora deve essere inviato ai delegati nazionali.
- Ci aspettiamo di ricevere gli **evaluation reports** dopo la metà di gennaio 2015.
- **Cross fingers** ☺

The Context of the EINFRA-1-2014 Call

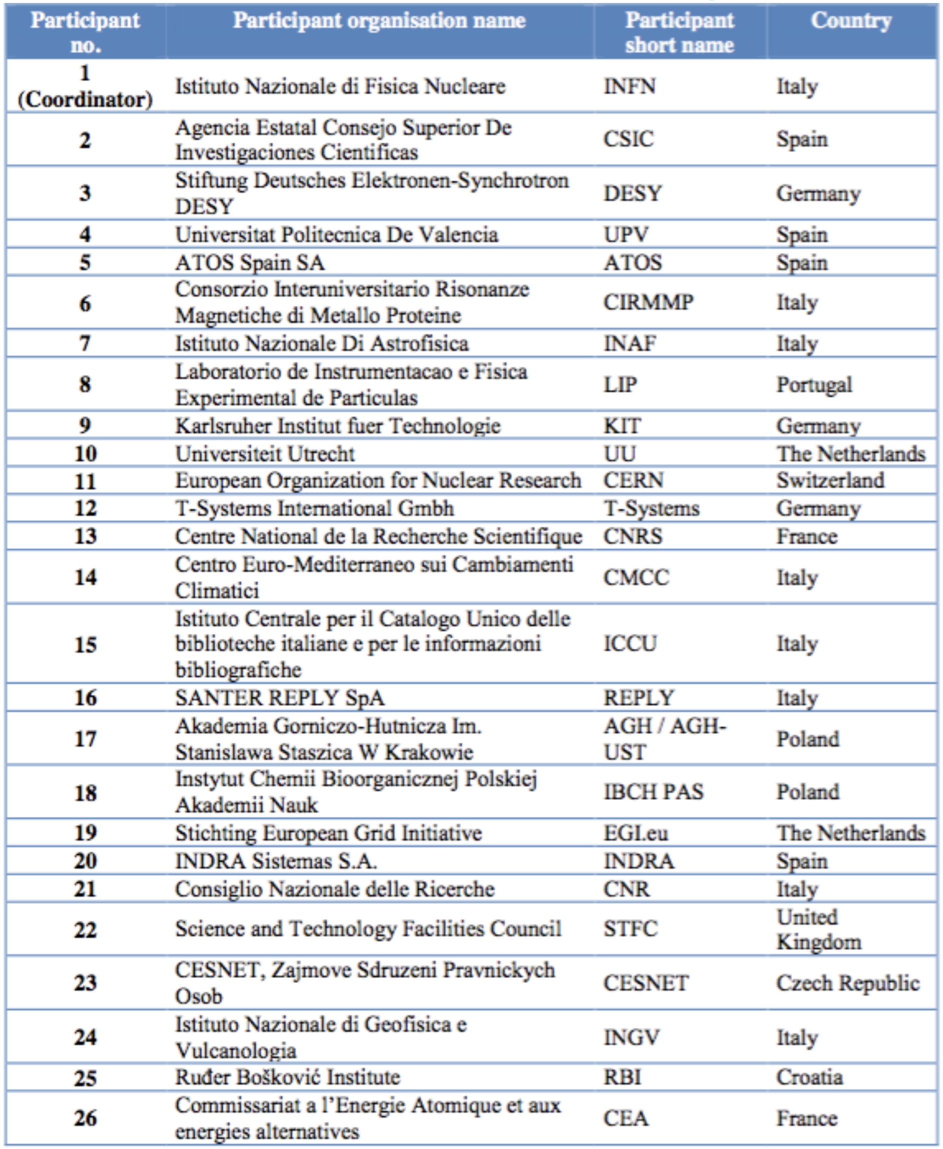
- **EINFRA-1-2014**, “Managing, preserving and computing with big research data”
  - Research and Innovation Action (RIA)
- **What**: “Development and deployment of integrated, secure, permanent, on-demand service-driven, privacy-compliant and sustainable e-infrastructures incorporating advanced computing resources and software.”
- **Why**: “increase the capacity to manage, store and analyze extremely large, heterogeneous and complex datasets[1], including text mining of large corpora.”
- **How**: “[P]rovide services cutting across a wide-range of scientific communities and addressing a diversity of computational requirements, legal constraints and requirements, system and service architectures, formats, types, vocabularies and legacy practices of scientific communities that generate, analyse and use the data.”

[1]: Research data include large datasets collected, developed or generated for/by research, integration of small distributed datasets, as well as data not originally collected for research, which may include environmental, social and humanities data.

EINFRA-1-2014: Items 4-5

- **Item 4:**
  - Large scale virtualization of data/compute centre resources to achieve on-demand compute capacities, improve flexibility for data analysis and avoid unnecessary costly large data transfers.
- **Item 5:**
  - Development and adoption of a standards-based computing platform (with open software stack) that can be deployed on different hardware and e-infrastructures (such as clouds providing infrastructure-as-a-service (IaaS), HPC, grid infrastructures…) to abstract application development and execution from available (possibly remote) computing systems. This platform should be capable of federating multiple commercial and/or public cloud resources or services and deliver Platform-as-a-Service (PaaS) adapted to the scientific community with a short learning curve. Adequate coordination and interoperability with existing e-infrastructures (including GÉANT, EGI, PRACE and others) is recommended.

Info di progetto



Participant no.	Participant organisation name	Participant short name	Country
1 (Coordinator)	Istituto Nazionale di Fisica Nucleare	INFN	Italy
2	Agencia Estatal Consejo Superior De Investigaciones Cientificas	CSIC	Spain
3	Stiftung Deutsches Elektronen-Synchrotron DESY	DESY	Germany
4	Universitat Politecnica De Valencia	UPV	Spain
5	ATOS Spain SA	ATOS	Spain
6	Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine	CIRMMP	Italy
7	Istituto Nazionale Di Astrofisica	INAF	Italy
8	Laboratorio de Instrumentacao e Fisica Experimental de Particulas	LIP	Portugal
9	Karlsruher Institut fuer Technologie	KIT	Germany
10	Universiteit Utrecht	UU	The Netherlands
11	European Organization for Nuclear Research	CERN	Switzerland
12	T-Systems International Gmbh	T-Systems	Germany
13	Centre National de la Recherche Scientifique	CNRS	France
14	Centro Euro-Mediterraneo sui Cambiamenti Climatici	CMCC	Italy
15	Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche	ICCU	Italy
16	SANTER REPLY SpA	REPLY	Italy
17	Akademia Gorniczo-Hutnicza Im. Stanislawa Staszica W Krakowie	AGH / AGH-UST	Poland
18	Instytut Chemii Bioorganicznej Polskiej Akademii Nauk	IBCH PAS	Poland
19	Stichting European Grid Initiative	EGI.eu	The Netherlands
20	INDRA Sistemas S.A.	INDRA	Spain
21	Consiglio Nazionale delle Ricerche	CNR	Italy
22	Science and Technology Facilities Council	STFC	United Kingdom
23	CESNET, Zajmove Sdruzeni Pravnickych Osob	CESNET	Czech Republic
24	Istituto Nazionale di Geofisica e Vulcanologia	INGV	Italy
25	Ruđer Bošković Institute	RBI	Croatia
26	Commissariat a l'Energie Atomique et aux energies alternatives	CEA	France

- 26 partner
- Durata del progetto: 30 mesi
- Budget totale: €11.138.114
- Budget INFN: €2.080.614 (18,7%)
- PM totali: 1580 (52,7 FTE)
- PM INFN: 315 (10,5 FTE)
- Sezioni INFN coinvolte: CNAF, Bari, Catania, Padova, Torino

Lettere di supporto

- Interesse esplicito a INDIGO (attraverso Letters of Support) è stato espresso da:
  1. Fernando Ballestero, Deputy Director General for International and European Affairs, **Ministry of Economy and Competitiveness of Spain**
  2. Paulo Pereira, member of the Board of Directors, **Portuguese Science Foundation (FCT)**
  3. Prof. Dorte Olesen, chair of the **GEANT Assembly**
  4. Dr. Laurent Romary, dr. Tobias Blanke and dr. Conny Kristel, directors of **DARIAH-EU** (ESFRI project)
  5. Dr. Paolo Favali, coordinator of **EMSO** (ESFRI project)
  6. Prof . Dave Stuart, dr. Susan Daenke, director and coordinator of **INSTRUCT** (ESFRI project)
  7. Sylvie Joussaume, chair of ENES Scientific Board and Coordinator of **IS−ENES2** project
  8. Prof. Jesus Carretero, **COST Action IC1305 (NESUS) Chair**, University Carlos III, Madrid
  9. Dean N. Williams, PI, **Earth Systems Grid Federation (ESGF)**, Project Leader for Analytics and Informatics Management System
  10. Luis Martí-Bonmatí, Director of the Biomedical Imaging Databank of the Valencia Region (BIMCV), **EuroBioImaging** ESFRI node
  11. Dr. Roberto Bilbao, director of Basque Biobank – **Basque Foundation for Health Innovation and Research (BIOEF)**
  12. Sanzio Bassini, Director of Supercomputing, Applications & Innovation Department at **CINECA**, hosting the Italian PRACE Tier-0
  13. Dr. Ian Foster, director of the Computation Inst. at the **Univ. of Chicago and Argonne National Lab.**
  14. James Taylor Ralph S. O'Connor Associate professor of Biology, Associate professor of Computer Science at **Johns Hopkins University for the Galaxy project**
  15. Prof. Andrew Lonie, head of Life Science Computation Centre, **Victorian Life Sciences Computing.**

Che cosa vogliamo fare (high-level)

- INDIGO aims at **developing a data/computing platform** targeted at scientific communities, deployable on multiple hardware, and provisioned over hybrid (private or public) e-infrastructures.
- In INDIGO, **key European developers, resource providers, e-infrastructures and scientific communities have joined** to ensure the successful exploitation and sustainability of the project outcome.
- A key strength and benefit of the INDIGO consortium, for both the scientific and industrial sectors, is the potential to **create a new sustainable Cloud competence in Europe for PaaS**, similar to what OpenNebula or OpenStack have done for IaaS.

Gli obiettivi generali

- **Objective 1**: Development of a Platform based on open source software, without restrictions on the e-Infrastructure to be accessed (public or commercial, GRID/Cloud/HPC) or its underlying software.
- **Objective 2**: Providing the interface between e-Infrastructures and Platforms.
- **Objective 3**: Provide high-level access to the platform services in the form of science gateways and access libraries.
- **Objective 4**: Objective 4: Streamline the adoption of the software products.

I principi guida

- **Develop an open framework**, based on open source software, and without initial restrictions on the e-Infrastructure to be accessed (public or commercial, GRID/Cloud/HPC) or its underlying software (OpenStack or OpenNebula). The consortium will take into consideration the medium and long-term exploitation and sustainability in this framework.

- **Exploit existing solutions**, learning, re-using and extending them according to user requirements, and having in mind the expected evolution of technology. INDIGO will also keep a direct contact with e-Infrastructure providers, to assure a successful deployment of the INDIGO Platform and to guarantee an adequate level of support.

- Ensure that the framework offered to final users as well as to developers will have a **low learning curve**, while providing a clear added value. In particular, existing software suites popular in different research communities, and offering already a rich environment for large data processing, like ROOT, OCTAVE/ MATLAB, MATHEMATICA or R-STUDIO, will be supported and offered in a transparent way to the final user, while running on powerful remote e-Infrastructure resources.

Un caso d'uso

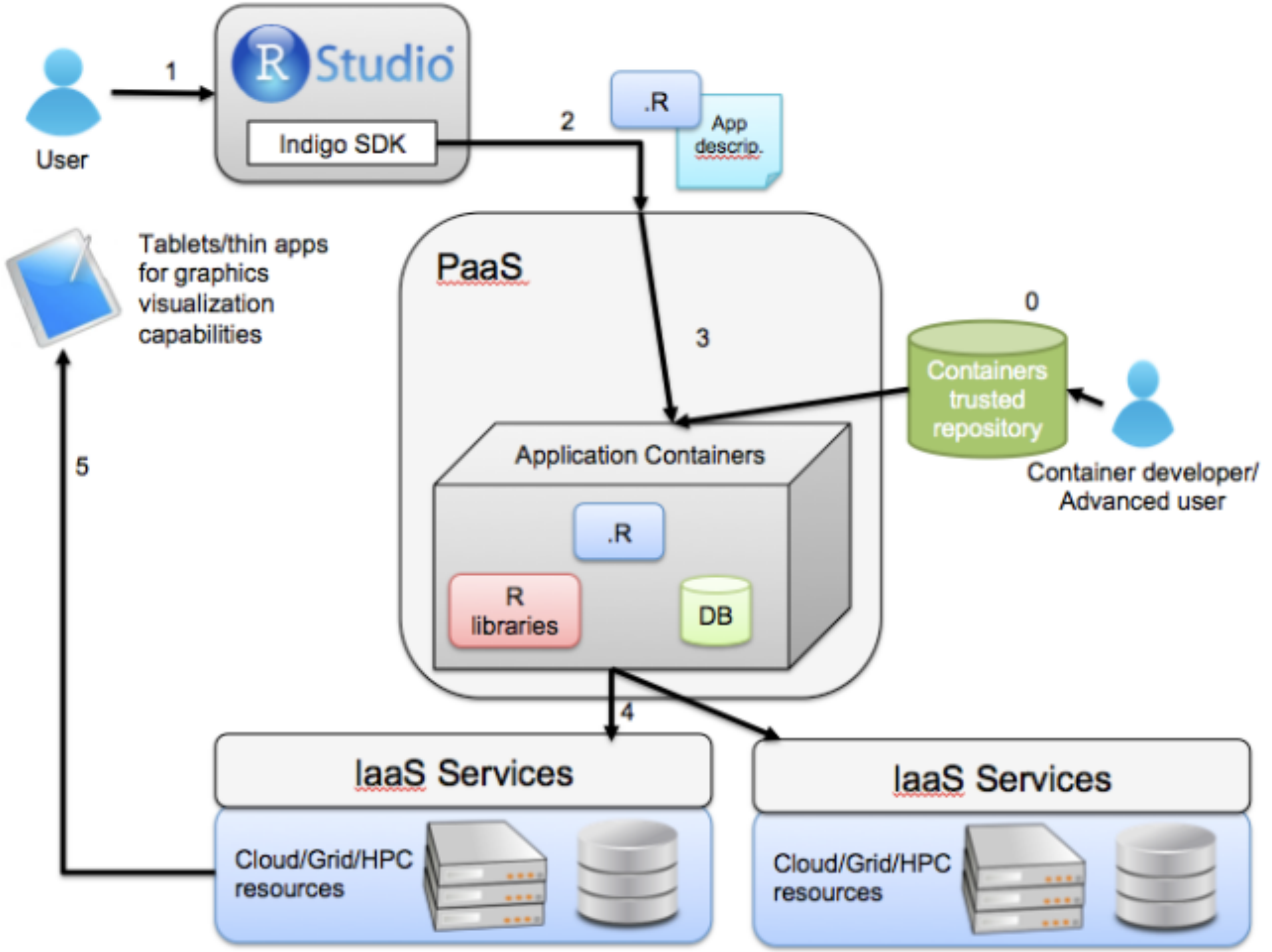


Figure 1: use case of supporting R-Studio through INDIGO

Work Packages (WP)

- Sono divisi tra WP di tipo Network Activity (NA), Service Activity (SA) e Joint Research Activity (JRA).
  - **WP1 (NA)**: admin and financial management, project quality assurance, global oversight.
  - **WP2 (NA)**: this is where Research Communities express requirements, provide feedback, review deployed services. Includes dissemination and communication activities.
  - **WP3 (SA)**: software management, deployment of pilot services.
  - **WP4 (JRA)**: resource virtualization.
  - **WP5 (JRA)**: PaaS framework.
  - **WP6 (JRA)**: APIs and portals, data-driven workflows.

Connessione tra WP

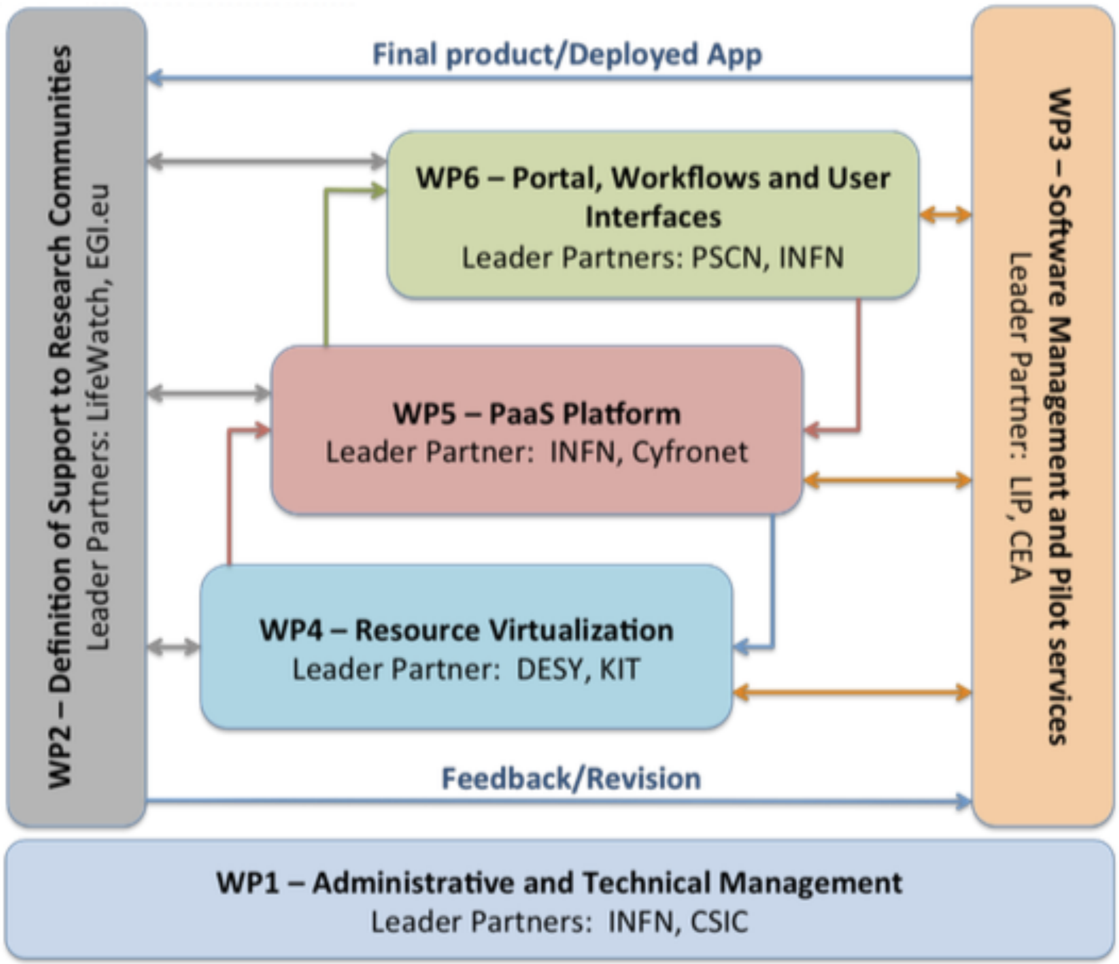

Final product/Deployed App
WP2 – Definition of Support to Research Communities
Leader Partners: LifeWatch, EGI.eu
WP6 – Portal, Workflows and User Interfaces
Leader Partners: PSCN, INFN
WP5 – PaaS Platform
Leader Partner: INFN, Cyfronet
WP4 – Resource Virtualization
Leader Partner: DESY, KIT
WP3 – Software Management and Pilot services
Leader Partner: LIP, CEA
Feedback/Revision
WP1 – Administrative and Technical Management
Leader Partners: INFN, CSIC

Figure 4: Diagram showing the interrelation among the Work Packages

WP Leaders/Deputies

- WP1: Davide Salomoni (INFN, PI) / Isabel Campos (CSIC)
- WP2: Jesus Marco (LifeWatch) / Peter Solagna (EGI.eu)
- WP3: Jorge Gomes (LIP) / Zdenek Sustr (CESNET)
- WP4: Patrick Fuhrmann (DESY) / Marcus Hardt (KIT)
- WP5: Giacinto Donvito (INFN) / Lukasz Dutka (Cyfronet)
- WP6: Marcin Plociennik (PSNC) / Roberto Barbera (INFN)

Architettura globale

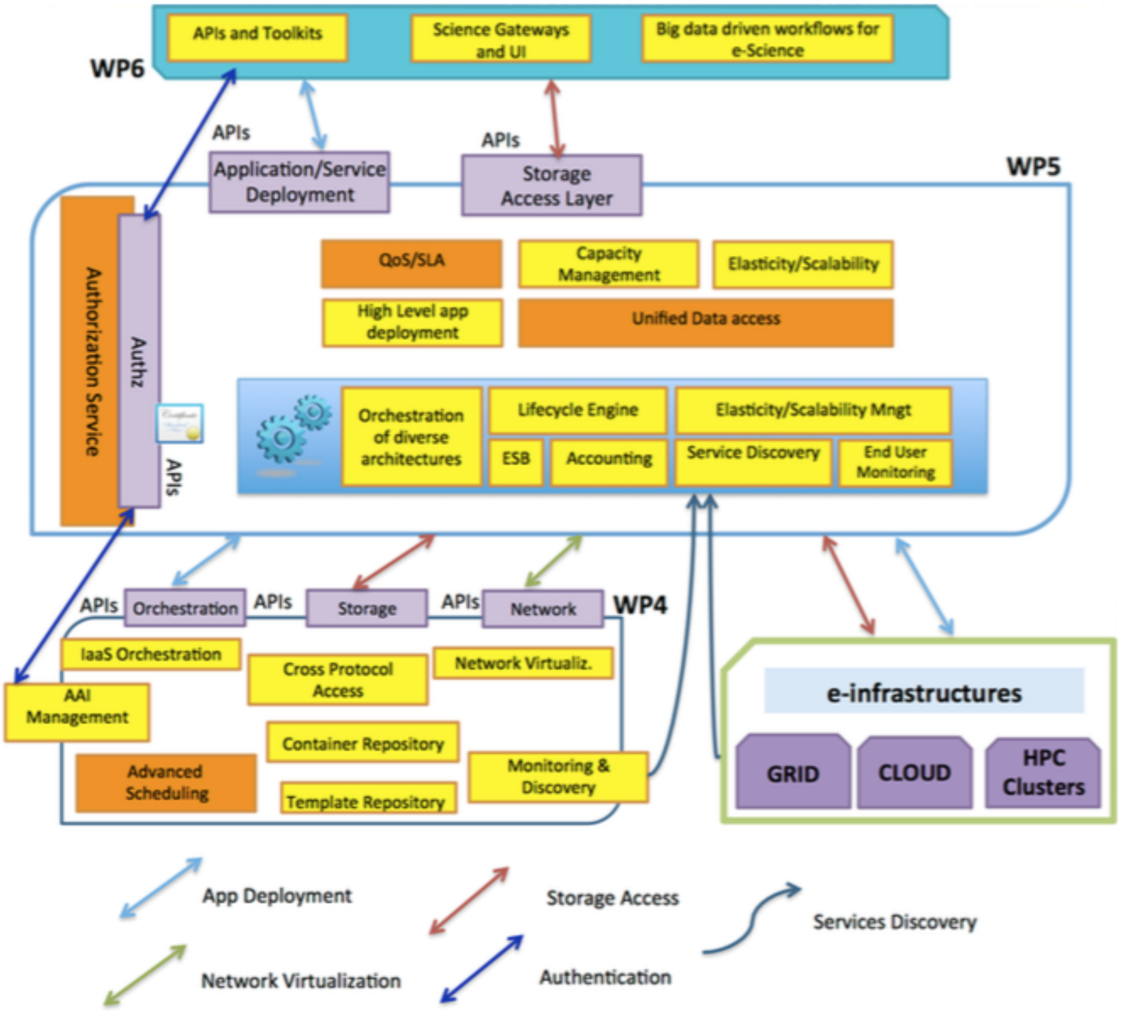
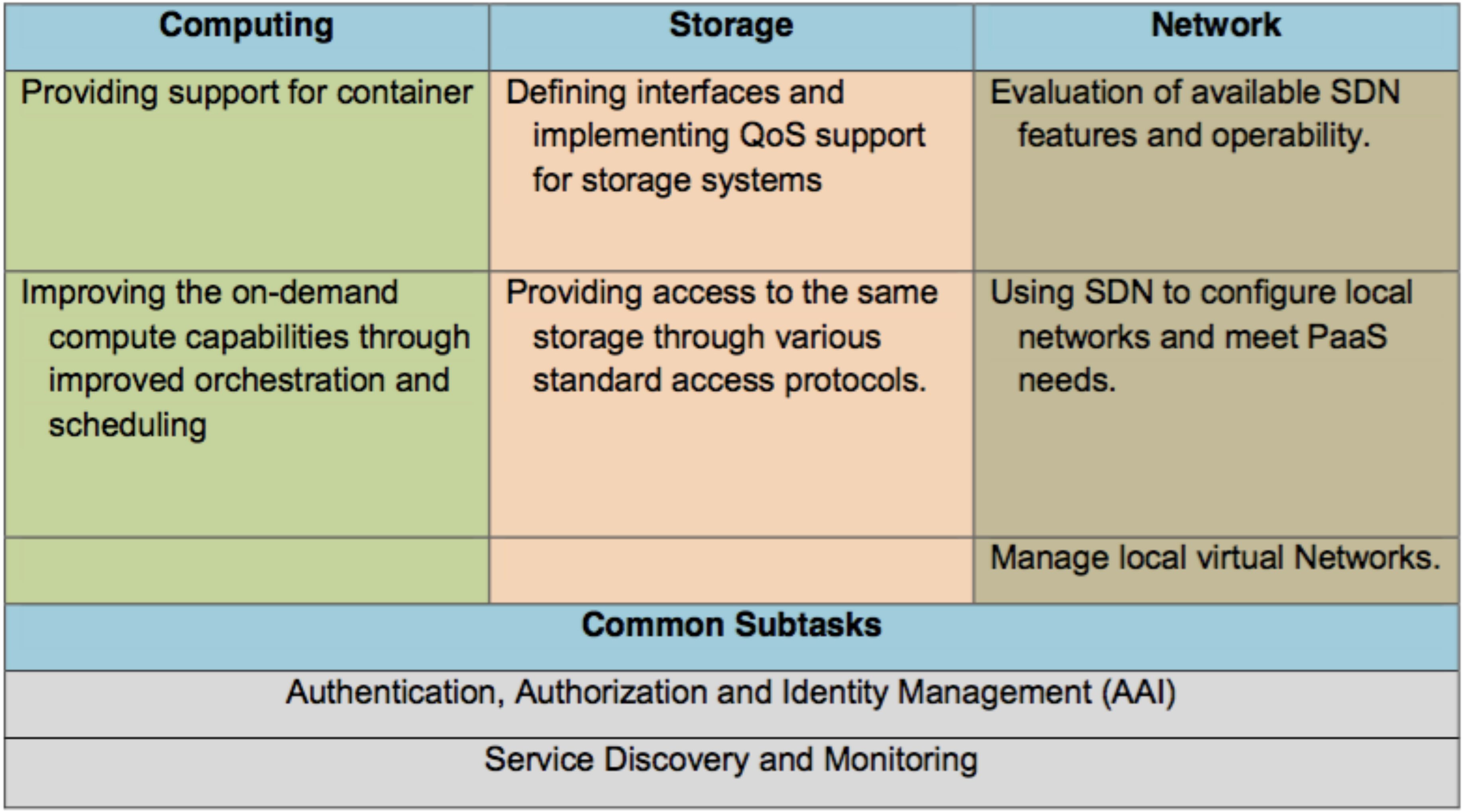


Figure 2: INDIGO global architecture. Color codes: Yellow: implementation based on already available solution to be improved/changed; Orange: newly implemented services.

WP4, resource virtualization



Computing	Storage	Network
Providing support for container	Defining interfaces and implementing QoS support for storage systems	Evaluation of available SDN features and operability.
Improving the on-demand compute capabilities through improved orchestration and scheduling	Providing access to the same storage through various standard access protocols.	Using SDN to configure local networks and meet PaaS needs.
		Manage local virtual Networks.
Common Subtasks		
Authentication, Authorization and Identity Management (AAI)		
Service Discovery and Monitoring		

Table 6: Breakdown of WP4 into specific and common tasks

WP4 on containers

- [WP4 will] develop or extend container support in OpenStack and OpenNebula providing that feature to higher level services.
- [WP4] will also offer trusted repositories simplifying the creation of new containers, while providing a reliable endorsement.
- [WP4] will extend the relevant IaaS standard interfaces (e.g. OCCI, EC2 or DMTF CIMI) to support the portable use across different cloud management frameworks avoiding vendor lock-in.

WP4 on storage virtualization

- **QoS support**: enable users to specify service quality policies for their data. In collaboration with RDA, we envision standardizing the associated terms and definitions, so users can expect the same quality of service regardless of the underlying implementation.

- **Cross-protocol support**: use cases often require storing files with one access protocol and subsequently accessing the same data with a different protocol. This requires enabling access to identical data via different protocols. Different storage systems assume various data organization and representation behind the access protocols their expose (e.g. object vs file storage). Appropriate translation technologies will have to be developed.

WP5, PaaS development

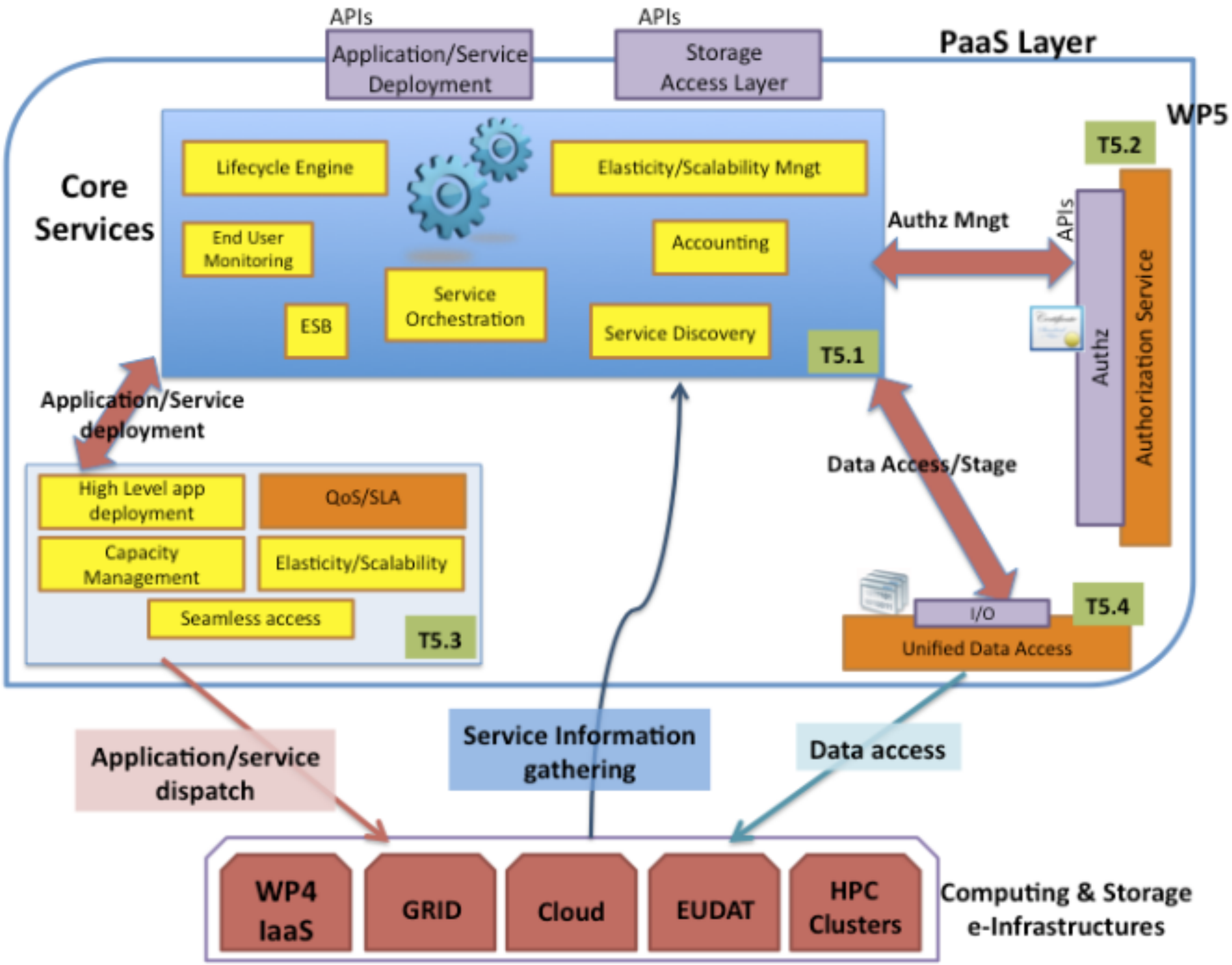


Figure 14: WP5 architecture

WP5 services (1)

- **Core PaaS**
  - An orchestration system that will receive standard descriptions of services or applications (e.g. TOSCA) and their interactions with other services or applications. It is a crucial component conveying interest of both research communities and industrial partners.
- **Authorization services**
  - Services and tools needed to enable a secure composition of services from multiple providers in support of scientific applications.

WP5 services (2)

- **Dispatcher service**
  - Deploy in a transparent way both services and applications in a distributed and heterogeneous environment made by different infrastructures (EGI Grid, EGI Fed Cloud, IaaS Cloud, Helix Nebula, PRACE, local HPC clusters, etc).
- **Unified data access layer**
  - This task will provide users and other INDIGO services a unified and transparent access to distributed storage. As a result, a set of APIs and services will be delivered to face problems of data orchestration, storage access optimization, remote data access. The task will focus on effective and reliable discovery of storage systems capabilities, understanding their data management policies, discovering Retention Time and Access Latency (e.g. out of meta-data or through SRM/CDMI interfaces).

WP6, Science gateways, workflows and toolkits

- **Libraries and toolkits**
  - Build a set of libraries and toolkits on the REST APIs developed by WP5. The aim of these libraries and toolkits is to simplify the development process and speed up the creation of science gateways and desktop and mobile applications.
- **Science gateways and mobile apps**
  - Main goal: develop Science Gateway software, based on the specific requirements of the WP2 communities, on top of the PaaS (WP5) that could be customized and deployed for the purpose of given community (in WP3).
- **Support for big data driven workflows for e-Science**
  - Based on the use cases gathered in WP2, this task will provide dynamic support for scientific workflows management according to a “Workflow as a Service” (WaaS) model. The proper workflow engines will be selected taking into account user needs and requirements. With regard to the different application scenarios, this task will provide workflow services able to seamlessly orchestrate workflows in Cloud, Grid and HPC environments. Input and output data could be stored on Grid, Cloud, local, external storage resources and will be accessed through standards interfaces (e.g. SAGA, OCCI, CDMI).

Governance

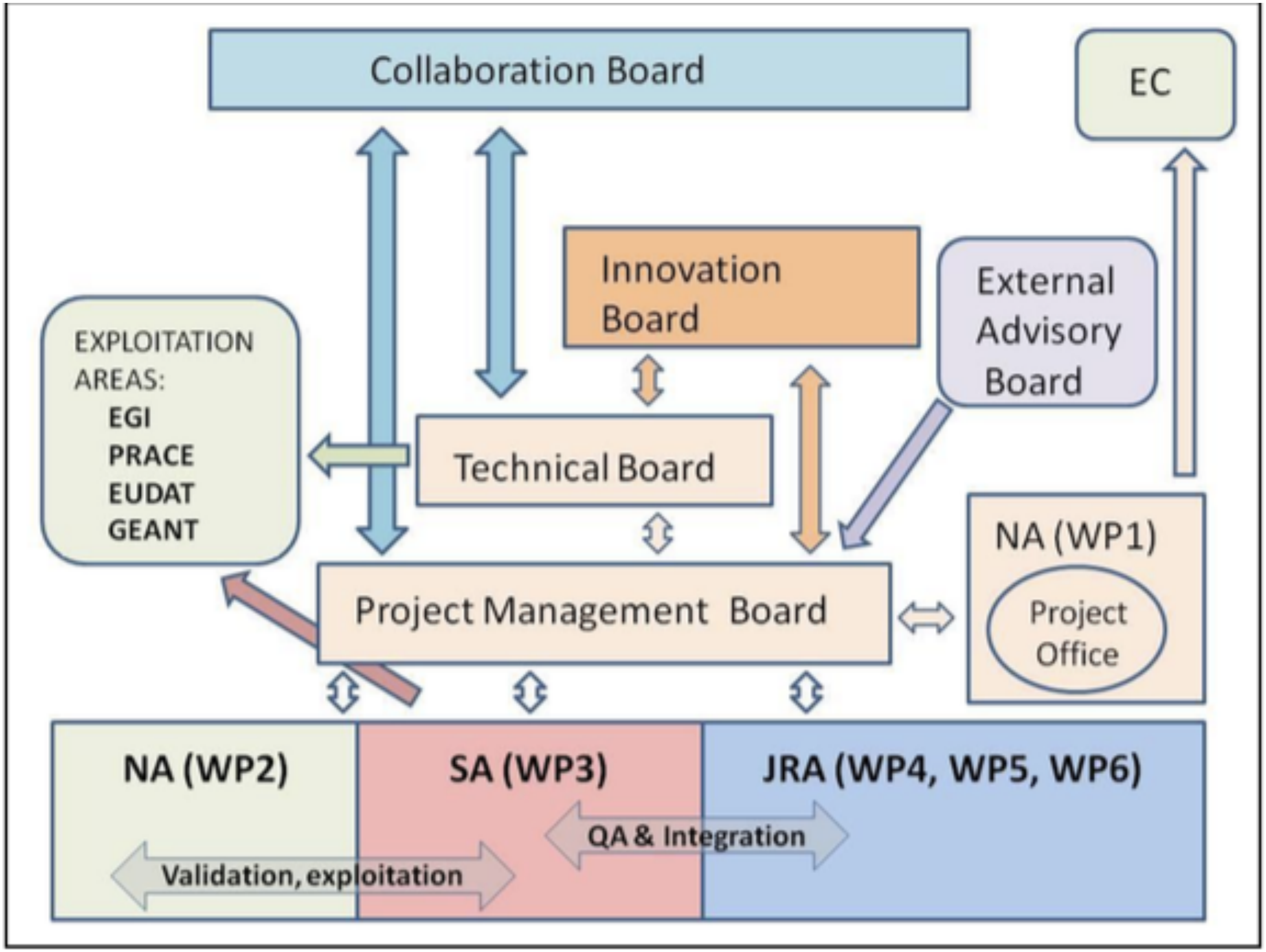


Figure 16 INDIGO Governance structure

I tempi

- **Se INDIGO viene approvato**, assumiamo di riceverne notizia ufficiale entro la fine di gennaio.
- La Commissione Europea stabilisce poi massimo 3 mesi per la firma del Grant Agreement.
- Potenzialmente dunque potremmo avere il kick-off di INDIGO tentativamente all'inizio di aprile (presumbilmente in Italia).
  - Vedremo comunque che cosa faranno anche gli altri progetti approvati nella call EINFRA-1. Una concertazione di tempi e (ove applicabile) obiettivi potrebbe essere strategicamente molto utile.
- Nel caso in cui sia INDIGO sia EGI-Engage venissero approvati, un "public event" post kick-off potrebbe essere organizzato a seguire la EGI TC prevista per fine maggio 2015.

Ringraziamenti

- La proposta INDIGO è stata completata in tempi molto rapidi, con la maggior parte del lavoro compiuto sostanzialmente durante il periodo estivo. Non è possibile ringraziare qui individualmente tutte le persone dell'INFN che hanno contribuito al progetto perché sono troppe. **Qualunque sarà l'esito della valutazione, comunque grazie!**
- Un **ringraziamento speciale** va a:
  - Nando Ferroni, Antonio Zoccoli, Luciano Gaido, Donatella Lucchesi, Veronica Valsecchi, Dario Menasce, Giacinto Donvito, Roberto Barbera, Gaetano Maron.