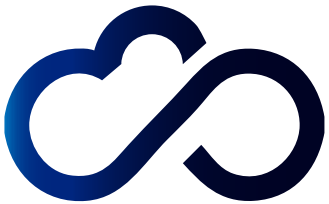


INDIGO DataCloud

Giacinto DONVITO
INFN-Bari

Outline

- EINFRA-1 Call
- INDIGO Consortium
- Community involved
- Gap Analysis
- Goals and Technology
- Conclusions

EINFRA-1-2014: Items 4-5 (specifically addressed by INDIGO)

- 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.

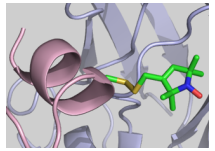
INDIGO-Datacloud

INtegrating **D**istributed data **I**nfrastructures for **G**lobal Expl**O**itation

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

Communities involved

- **Biological and medical science**
  - Biological, molecular and medical imaging, life science research applied to medicine, agriculture, bio-industries and society, structural biology.



- **Social sciences, arts and humanities**
  - Georeferencing (e.g. of current or historical maps), cultural heritage, smart sensors.

- **Environmental and earth science**
  - Biodiversity and ecosystem research, interactions between geosphere, biosphere and hydrosphere, earth system modeling.

- **Physical sciences**
  - Astrophysics, theoretical and experimental research in physics.

Gap Analysis

- Support **federated identities** and provide privacy and distributed authorization in open **Cloud platforms**
- **Performance issues** limiting massive adoption of virtualized Cloud resources in **large data centers**.
- **Orchestrate** and federate **Cloud** [public or private], **Grid** and **HPC** resources
- The barriers that **limit the adoption** of true **PaaS** solutions, such as the use of custom, **non-interoperable interfaces** and the limited availability of APIs for technology-independent storage access
- The lack of **flexible data sharing** between groups' members and the difficulty in obtaining **easy access to data** generated by collaborating users working with different infrastructures or sites

Gap Analysis

- **Static allocation** and partitioning of both storage and computing **resources** in data centers
- Avoid software and **vendor lock-in**
- Exploit **specialized hardware**, such as GPUs or low-latency interconnections
- Manage dynamic and **complex workflows** for scientific **data analysis**
- Provide **APIs** to exploit the capabilities of the infrastructure and write applications, **customizable portals and mobile views**
- The current **inflexible ways** of **distributing** and deploying applications.

Gap Analysis

- The lack of an open and solid platform that permits to exploit distributed computing and storage resources through **transparent network interconnections**.
- The difficulty of modifying existing HTC (High Throughput Computing) or HPC (High Performance Computing) applications so that they can exploit distributed Cloud resources and flexible workflows.

Use Cases

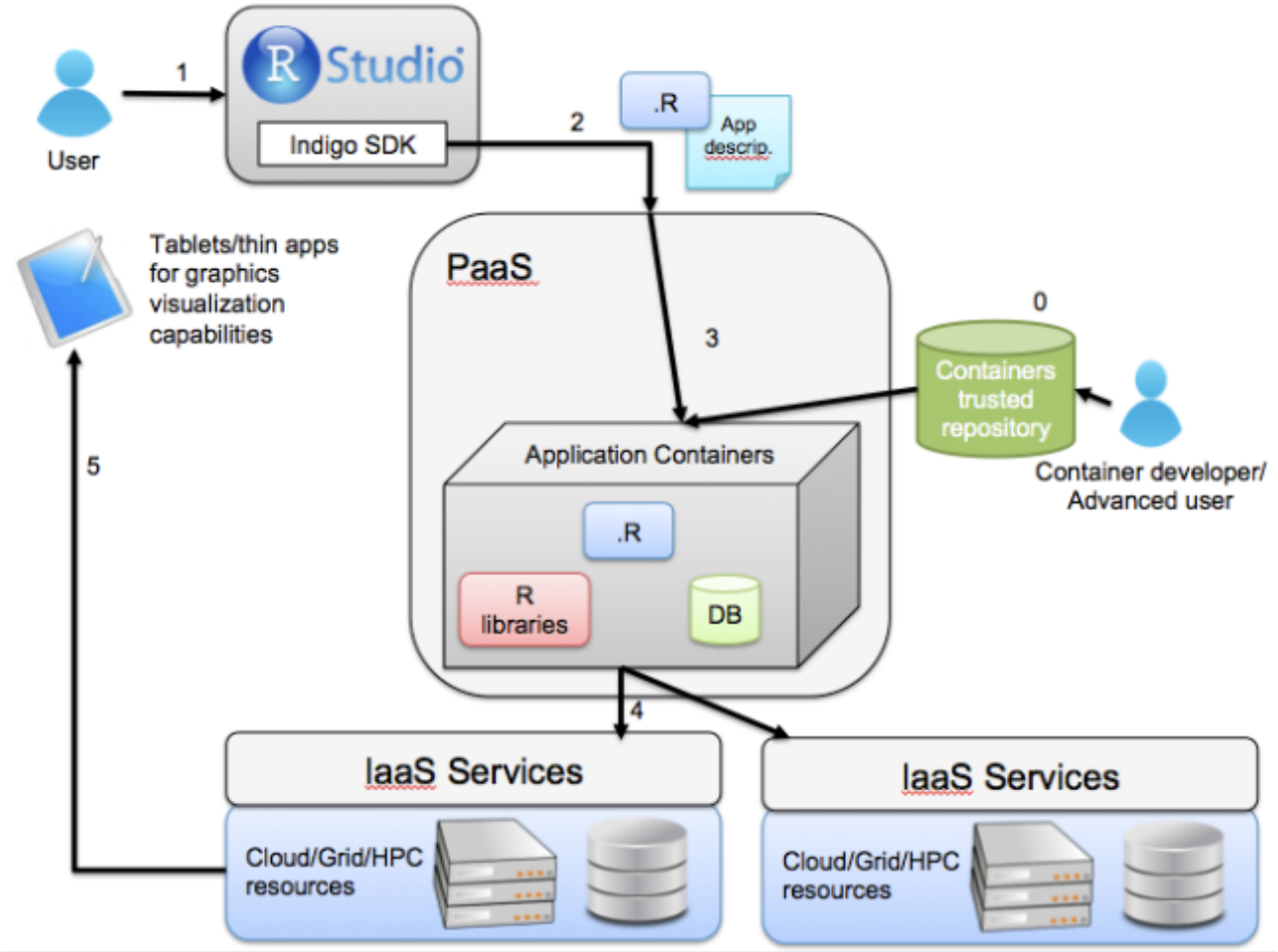
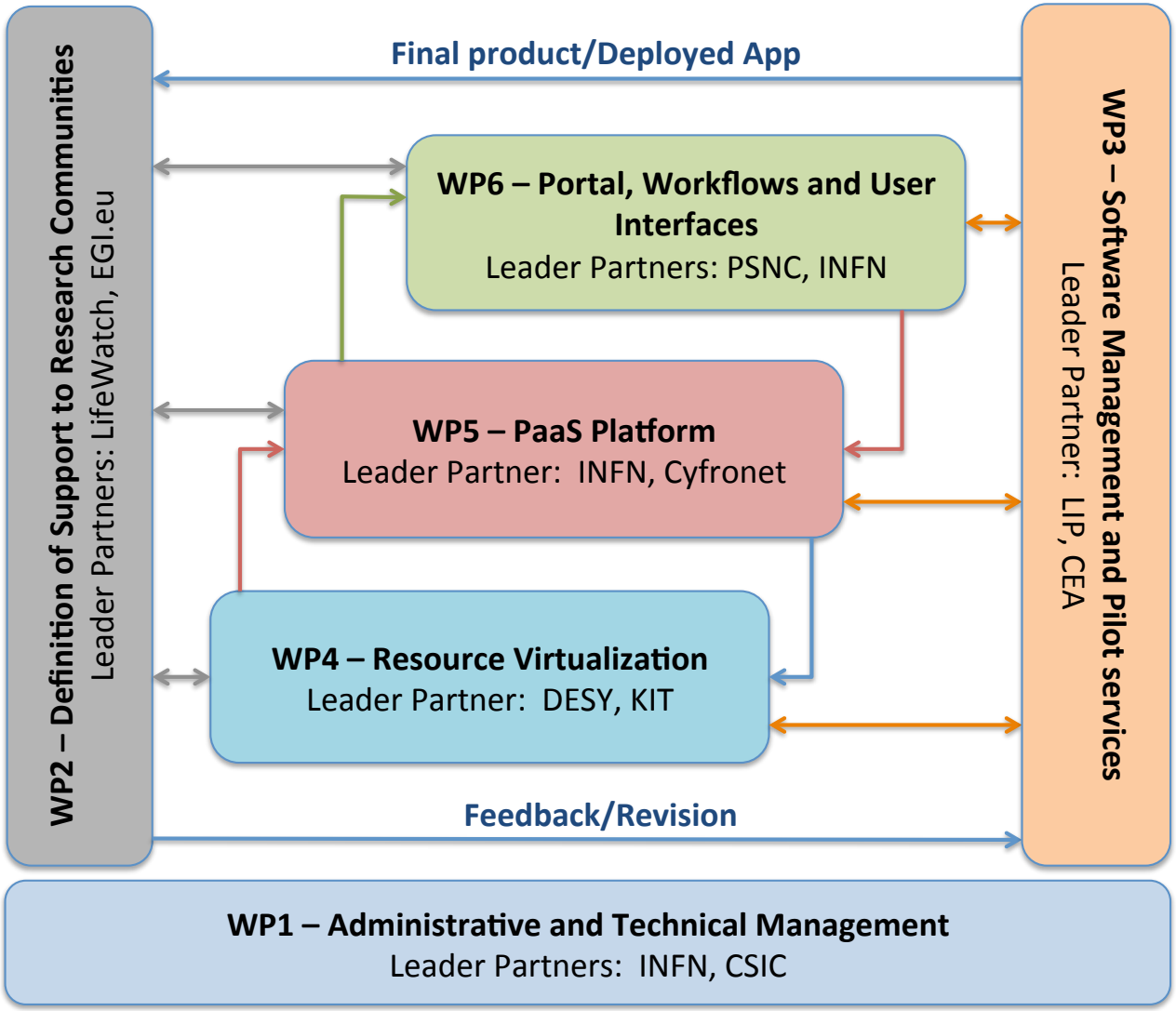


Figure 1: use case of supporting R-Studio through INDIGO

Work Packages



INDIGO approach

- Based on **Open Source** solutions
- widely **supported** by big communities
- whenever possible exploit **general solutions** instead of specific tools/services
  - or put effort in **increasing the generality** of tools developed in a given community
  - this will be important for **sustainability** of the architecture
- ensure that the framework offered to final users, as well as to developers, will have a **low learning curve**
  - **existing software suites** like ROOT, OCTAVE/MATLAB, MATHEMATICA or R-STUDIO, **will be supported** and offered in a transparent way

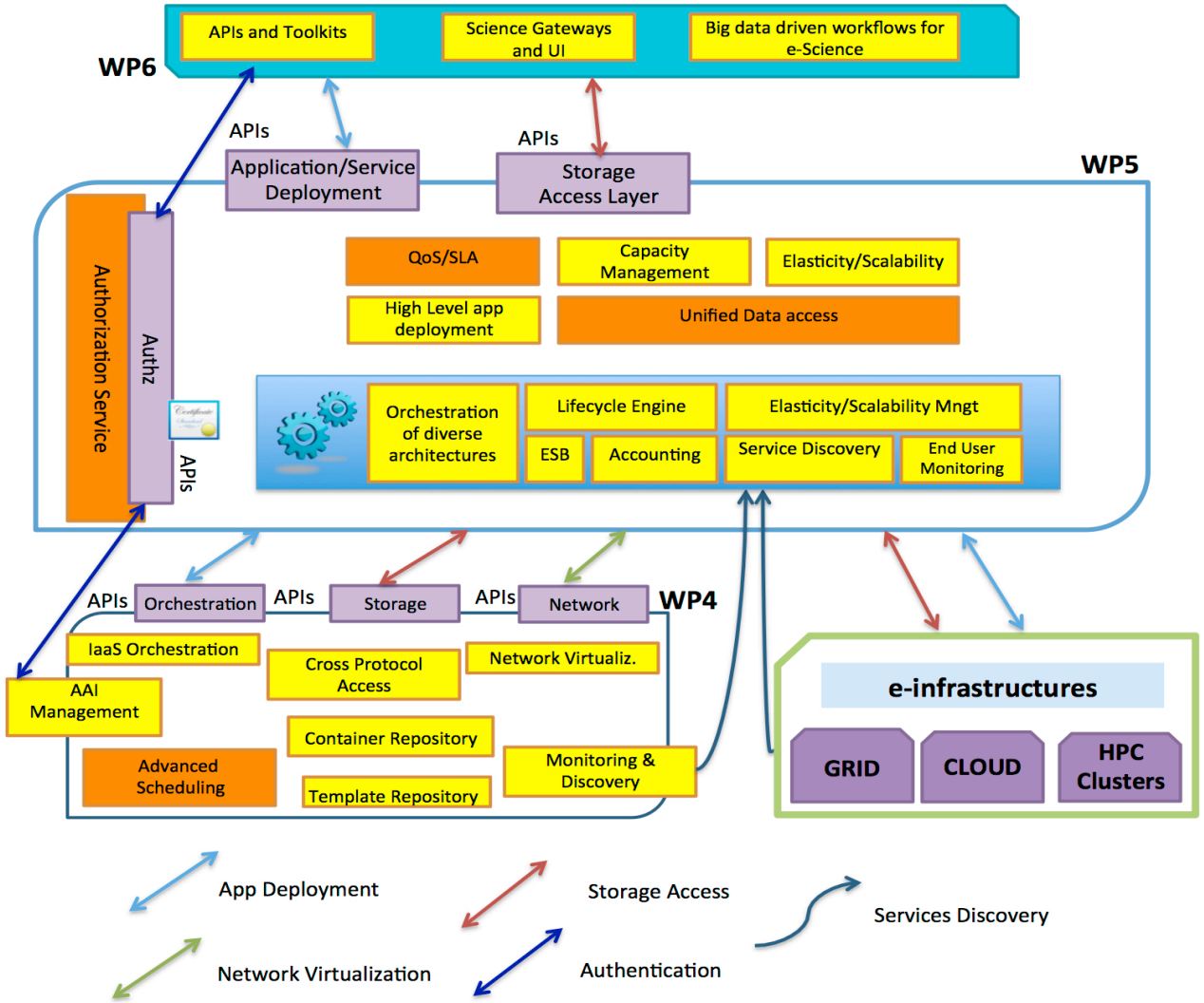
INDIGO Global architecture

INDIGO global architecture.

Color codes:

Yellow: implementation based on already available solution to be improved/ changed;

Orange: Completely new services to be implemented



WP3: Software Management and Pilot Services

WP3: Software Management and Pilot Services

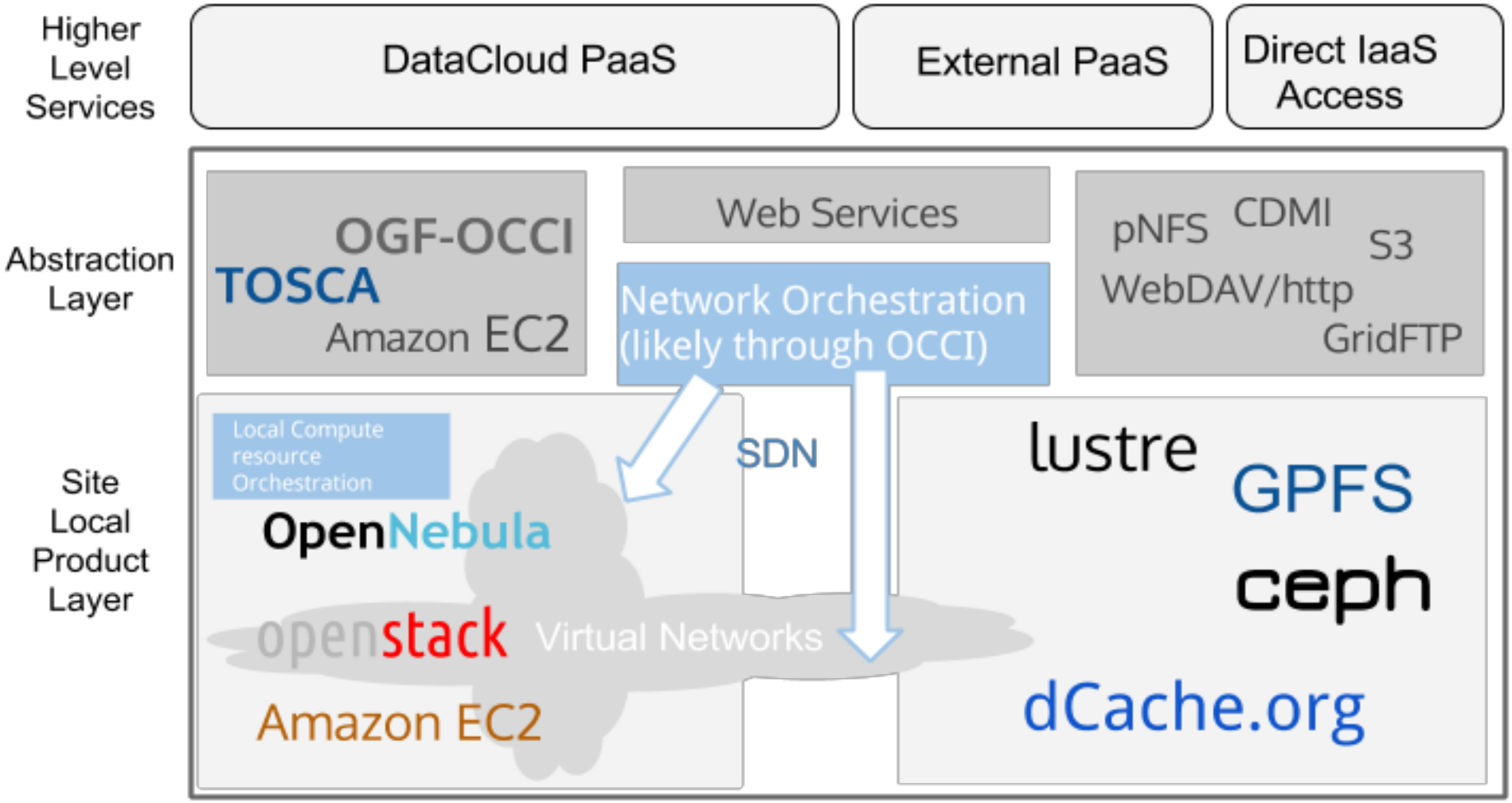
- T3.1 will deal with the **software quality assurance**, compiling and enforcing the necessary quality criteria, indicators, and tests necessary to ensure high quality software components ready for production.
- T3.2 will make the certified software components available as a set of coherent **high quality releases**, supported by an efficient maintenance process.
- T3.3 will provide the **pilot infrastructures** and services for integration and testing supporting the tasks T3.1 and T3.2 activities.
- T3.4 will interface with major production e-infrastructures, collect their feedback, requirements, and will enable a path **towards production exploitation**.

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: Resource Virtualization



WP4: Resource Virtualization

Common subtask:

- **AAI Management** for the virtualized computing cloud infrastructure
  - To ensure the integration of federated AAI technologies into OpenStack, OpenNebula, CEPH, dCache and other supported INDIGO products, allowing users to **access infrastructure** resources using their **home** or guest **IdP account**.
- **Service Discovery** and Monitoring
  - To extend existing **local-site monitoring services** for all INDIGO products to provide to higher-level services of WP5 and WP6 with monitoring and accounting information through a query-API.

WP4: Resource Virtualization

Cloud Computing Virtualization

- Providing **support for containers**
  - as a portable and performant platform for the execution and deployment of applications, and by providing local site **orchestration features** (e.g. HEAT or OneFlow) that simplify the management of the lifecycle of IaaS: both containers and VMs.
- Improving the on-demand compute capabilities of data-centers by **improving compute orchestration and scheduling**
  - To improve the existing **cloud schedulers** in **OpenStack and OpenNebula** to include the support for postponing low priority workloads (by killing, preempting or stopping running containers or VMs) in order to allocate higher priority requests, thus enabling the **advanced scheduling policies**, optimizing the usage of the data center and improving its response to the users.

WP4: Resource Virtualization

- **Cloud Storage Virtualization**
- **QoS Support** in storage
  - will 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.
  - Evaluate/extend **available protocols** (e.g. CDMI, WebDAV, SRM) supporting the defined service levels.
- **Cross Protocol support** for storage solutions
  - 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.

WP4: Resource Virtualization

- **Network Virtualization**
- **Enhancing** the capabilities of **Local virtual networks**
  - features comparable to real-life physical networks, including a pre-defined complex topology of such a network or the presence of active elements such as switches or routers.
- Use of **SDNs** to set up **virtual networks** spanning **multiple sites**
  - 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.

WP5: PaaS Platform Development

- **PaaS architecture and implementation**
  - Design and implementation of a PaaS layer allowing scientific communities to exploit, in a powerful and high-level way, **several heterogeneous computing and data e-infrastructure** such as: IaaS Cloud, Helix Nebula, EGI Grid, EGI Federated Cloud, PRACE, HPC, EUDAT, etc. It will be possible to process **large amounts of data** and to exploit the efficient storage and preservation technologies and infrastructure already available in the European e-infrastructure, with the appropriate mechanisms to ensure security and privacy.
  - Implementation of an advanced PaaS Layer providing the following features:
    - *Transparency*
    - *Error management*
    - *Elasticity/SLA management*.

WP5: PaaS Platform Development

- **Security and Authorization**
  - Will focus on **authorization issues** in distributed, multi-tenant **cloud infrastructures**, leveraging and integrating with standard and already available authentication technologies (e.g., **X.509, SAML2, OpenID-Connect**).
  - user identities, enrollment and attribute management (e.g., **group membership management, role assignment**);
  - access **policy definition**, distribution and composition;
  - controlled **privilege delegation**;
  - **credential translation**.

WP5: PaaS Platform Development

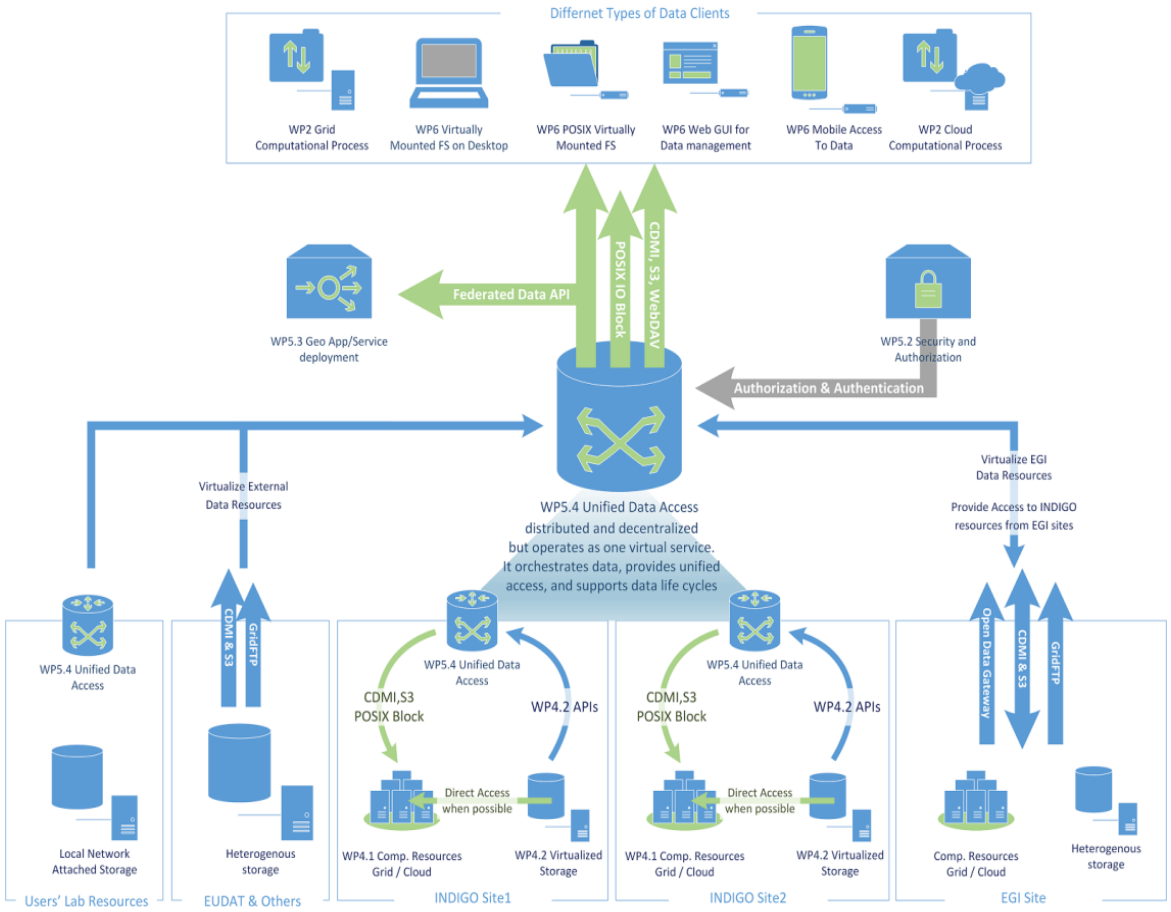
- **High-level geographical application/service deployment**
  - will implement a solution, **Geo-deployment Service**, to deploy in a transparent and powerful way both **services and applications** in a distributed and heterogeneous environment made by several different infrastructures (EGI Grid, EGI Fed Cloud, IaaS Cloud, Helix Nebula, PRACE, local HPC clusters, etc).
  - well **beyond the simple scheduling** of the application over grid or cloud environments because it will provide the capabilities to deploy a **wide range of diverse applications and services** with a powerful set of APIs that hide the complexity of the underlying infrastructures

WP5: PaaS Platform Development

- **Unified Data Access**
  - Unification
  - Federated data access
  - Interoperability and Open Data
  - Optimization and Data on the fly

WP5: PaaS Platform Development

- Unified Data Access



WP6: Science Gateways, Workflows and Toolkits

- **Develop Toolkits** (libraries) that will allow the platform usage from the level of the Scientific Gateways, desktop and mobile applications
- Provide and develop the **Open Source Mobile Application Toolkit** for the iOS, Android and WindowPhone platform that will be the base for development of the Mobile Apps.
- Provide the **User Friendly front end's**, that will prove the usability of the PaaS proposed:
  - Provide both a **general-purpose multi-domain Science Gateway** and customized examples for selected user communities/scenarios, that will make use of the proposed Toolkits, including **Data Analytics Gateways for e-Science**;

WP6: Science Gateways, Workflows and Toolkits

- Develop **example cross platform** native **Mobile Apps** for selected use cases, based on the Mobile App Toolkit;
- Manage the execution of **complex workflows using PaaS layers**;
- Support for both **interactive** and **batch** parallel data analytics **workflows**.

- Provide the dynamic scientific workflows services in a **Workflows as a Service** model.

- Provide workflow interfaces extensions for distributed and **parallel data analytics** on large volume of scientific, multidimensional data).

Conclusions

- INDIGO aims to **fill important gaps** in the field of **cloud computing** for e-Science
  - Enabling resource provider to **improve the efficiency** of their cloud infrastructures
  - Enabling users to exploit **available infrastructures** in an easier and efficient way
- INDIGO aims to develop **general purpose solutions**
  - That could be used not only by the user-communities already joining the project, but more widely from scientific communities that has similar needs
- INDIGO aims to build a **sustainable PaaS level** cloud solution for e-science based on widely used and supported technologies
- **WLCG** has a **long experience** in running very large distributed infrastructures in production for **scientific applications**
  - we are open to explore topics and technologies in which we could **fruitful co-operate**

Back-up slides

INDIGO Budget

- INDIGO will likely start in April 2015
- Duration:
  - 30 months
- The total budget is:
  - 11.1 M€

Part. No.	Participant short name	WP1	WP2	WP3	WP4	WP5	WP6	Total PMs / Participant
1	INFN	35	0	42	55	136	47	315
2	CSIC	30	30	29	30	30	0	149
3	DESY	0	0	0	60	30	0	90
4	UPV	0	15	0	12	45	0	72
5	ATOS	0	0	0	15	30	0	45
6	CIRMMP	0	30	0	0	0	0	30
7	INAF	0	15	0	0	15	0	30
8	LIP	0	0	39	40	0	0	79
9	KIT	0	0	0	42	11	11	64
10	U. Utrecht	0	24	0	0	0	0	24
11	CERN	0	0	5	10	20	0	35
12	T-SYSTEMS	0	0	0	0	30	0	30
13	CNRS	0	0	0	20	0	10	30
14	CMCC	0	12	0	0	0	27	39
15	ICCU	0	24	0	0	0	0	24
16	REPLY	0	0	5	10	45	0	60
17	AGH/AGH-UST	0	0	15	0	90	0	105
18	IBCH PAS	0	0	15	20	20	42	97
19	EGI.eu	0	34	26	0	0	0	60
20	INDRA	0	0	0	0	30	0	30
21	CNR	0	24	0	0	0	0	24
22	STFC	0	0	15	0	15	0	30
23	CESNET	0	0	13	28	13	0	54
24	INGV	0	24	0	0	0	0	24
25	RBI	0	30	0	0	0	0	30
26	CEA	0	0	10	0	0	0	10
	Total Person Months	65	262	214	342	560	137	1580