INDIGO - DataCloud

RIA-653549

INDIGO-DataCloud in the EOSC Context

INDIGO Summit 2017
9-12 May 2017, Catania

Better Software for Better Science.

Davide Salomoni, INFN-CNAF
INDIGO-DataCloud Project Coordinator

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

Acknowledgments First

INDIGO - DataCloud

- INDIGO is a project run by an outstanding set of collaborative, knowledgeable, and goal-oriented people. **Thanks and kudos to all of them**.

INFN Istituto Nazionale di Fisica Nucleare

Universiteit Utrecht

CESNET

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

PSNC

cea energie atomique • energies alternatives

cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici

AGH AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY

Reply

ICCU

DESY

CERM Firenze

egi

KIT Karlsruher Institut für Technologie

Science & Technology Facilities Council

RB

cnrs dépasser les frontières

LABORATÓRIO DE INSTRUMENTAÇÃO E FÍSICA EXPERIMENTAL DE PARTICULAS LIP

T · · Systems · · ·

CERN

Consiglio Nazionale delle Ricerche

indra

Atos

Istituto Nazionale di Geofisica e Vulcanologia

INAF ISTITUTO NAZIONALE DI ASTROFISICA NATIONAL INSTITUTE FOR ASTROPHYSICS

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

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.

INGV
egi
Rudjer Bošković Institute
Utrecht University
cmcc
Atos
T-Systems
indra
Reply
cnrs
Science & Technology Facilities Council
LIP
cea
INFN
CSIC
PSNC
CESNET
KIT

INDIGO-DataCloud's Vision

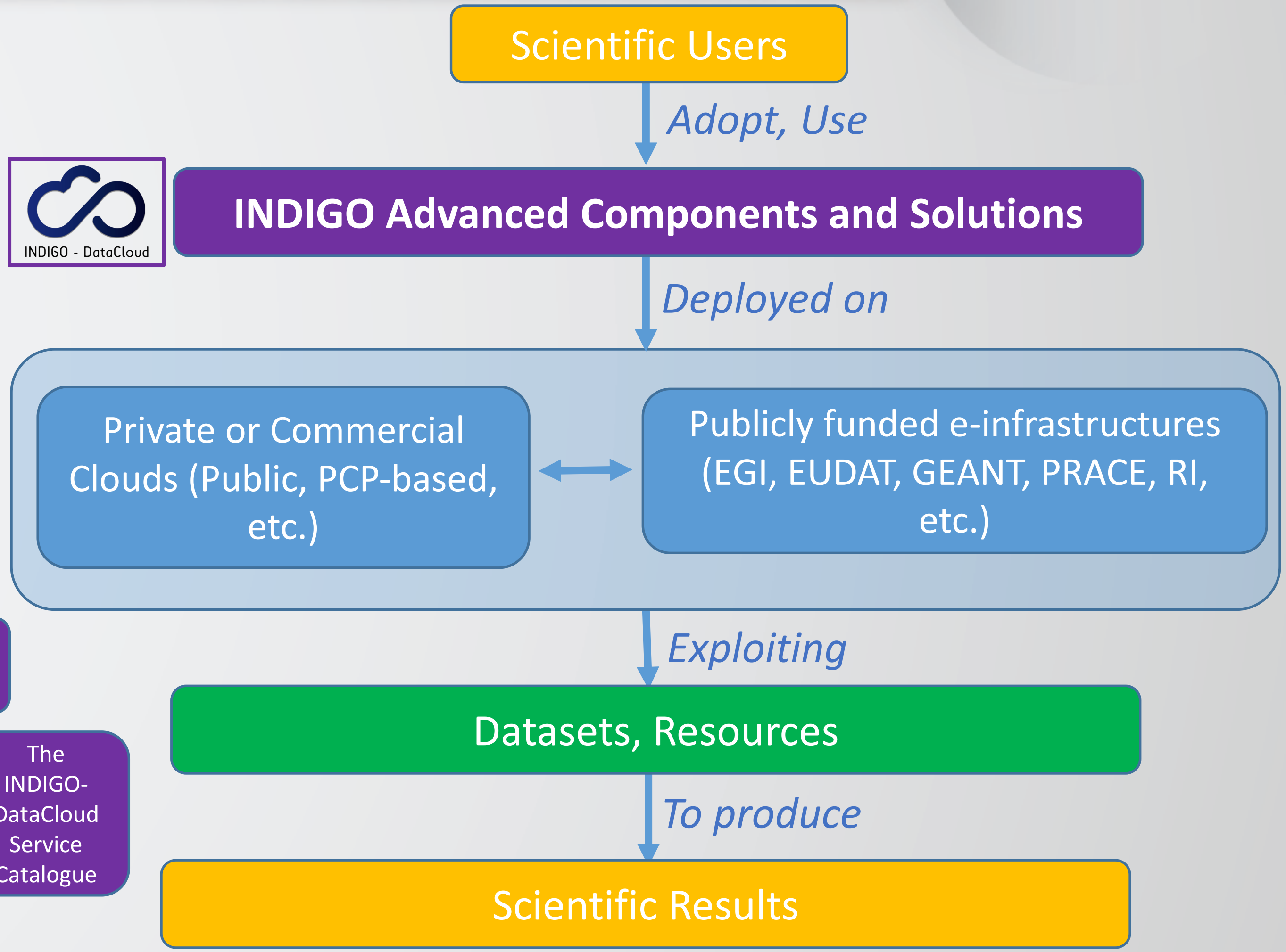
- INDIGO:
  1. **Develops open, interoperable solutions for scientific data.**
  2. **Supports open science** organizing the **European data space**.
  3. **Enables collaborations** across diverse scientific communities worldwide.
- INDIGO offers its
  - architecture,
  - analysis,
  - expertise
  - and software components
- as a **concrete step toward the definition and implementation of a European Open Science Cloud and Data Infrastructure**.

D1.8, General Architecture

D2.1 and D2.4, community requirements

INDIGO's 34 deliverables (so far)

The INDIGO-DataCloud Service Catalogue



Scientific Users

Adopt, Use

INDIGO Advanced Components and Solutions

Deployed on

Private or Commercial Clouds (Public, PCP-based, etc.)

Publicly funded e-infrastructures (EGI, EUDAT, GEANT, PRACE, RI, etc.)

Exploiting

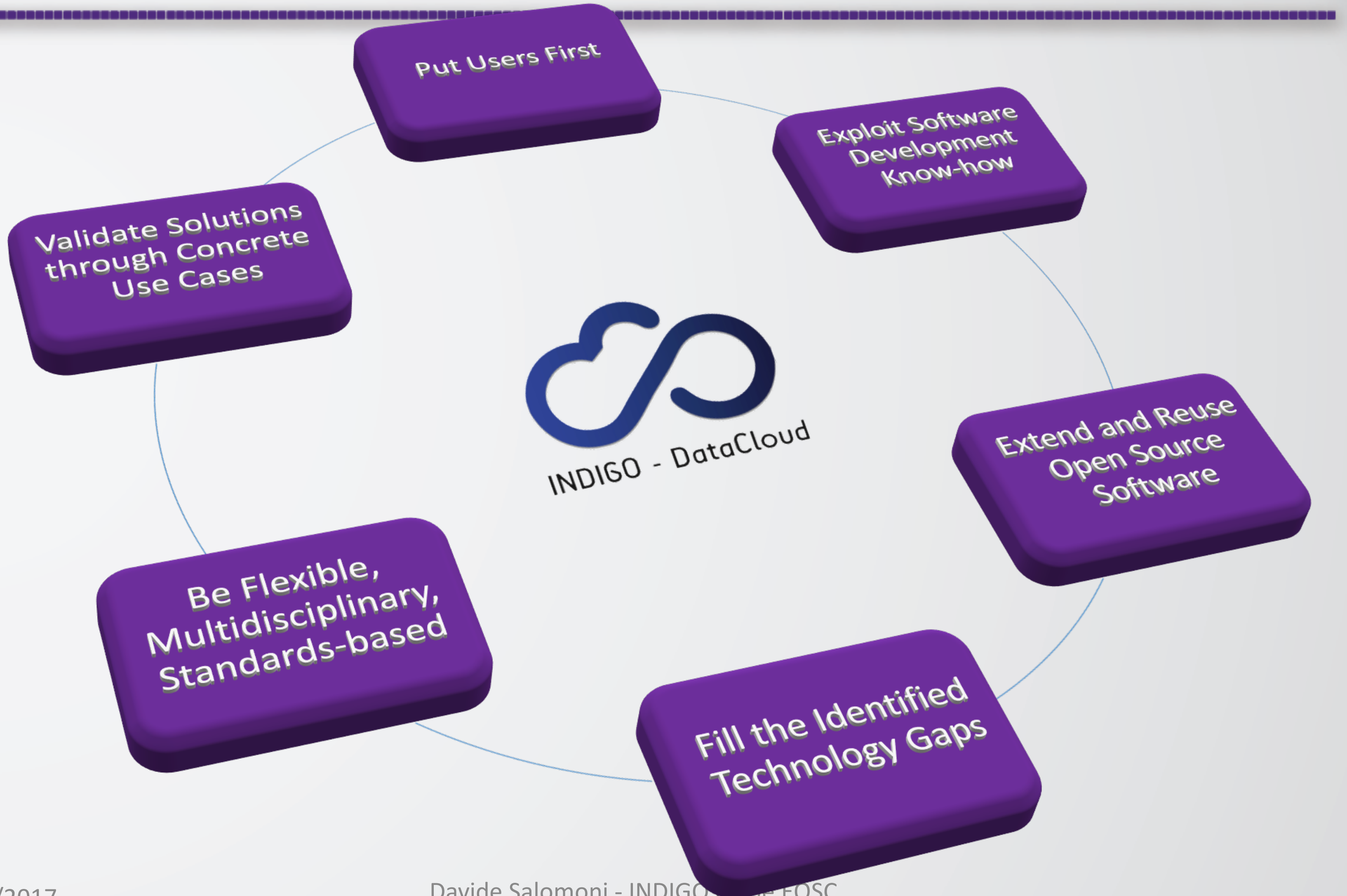
Datasets, Resources

To produce

Scientific Results

The INDIGO Foundations

INDIGO - DataCloud



Put Users First

Exploit Software Development Know-how

Extend and Reuse Open Source Software

Fill the Identified Technology Gaps

Be Flexible, Multidisciplinary, Standards-based

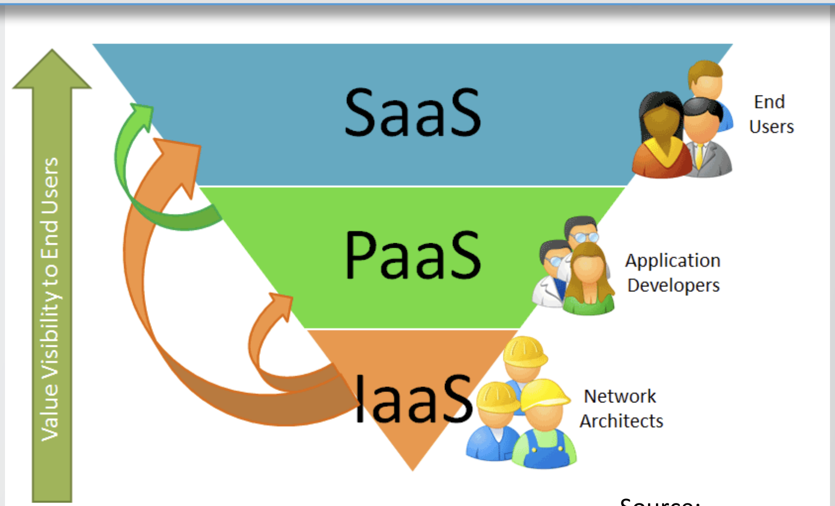
Validate Solutions through Concrete Use Cases

What did INDIGO originally want to address?

- Open **interoperation** / federation across (proprietary) Cloud infrastructures at the
  - IaaS,
  - PaaS,
  - and SaaS levels
- Managing **multitenancy**
  - At large scale…
  - … and in heterogeneous environments
- Handle dynamic and seamless **elasticity**
  - For both private and public clouds…
  - … for complex or infrequent requirements…
  - … through expressive and simple to use methods
- **Data management** in a Cloud environment
  - Tackling QoS, data replication, caching, transparent remote access

Addressing all of this should lead to:

- Interoperable PaaS/SaaS services addressing both public and private Cloud infrastructures.
- Porting of legacy applications to the Cloud.
- Increased focus on user-oriented, high-value solutions.



Source: https://goo.gl/cWZhKN

From the INDIGO-DataCloud proposal

[…] numerous areas are of interest to scientific communities where Cloud computing uptake is currently lacking, especially at the PaaS and SaaS levels.

The project therefore aims at developing tools and platforms based on open source solutions addressing scientific challenges in the Cloud computing, storage and network areas.

What INDIGO actually did

- INDIGO, driven by scientific communities, has been developing a **comprehensive open source Cloud architecture**, which provides **many new functionalities previously unavailable in open source and in some cases also in proprietary Cloud offerings.**
- These functionalities abstract from underlying IaaS technologies through the consistent use of both de jure and de facto standards. This allows **interoperability with hybrid (public/private) infrastructures.**
- After beta testing and demos shown as early as November 2015 (at the EGI Community Forum), **we released our first major software release (MidnightBlue) in August 2016, 9 software updates in the following months, and our second and final major release (ElectricIndigo) in April 2017.**

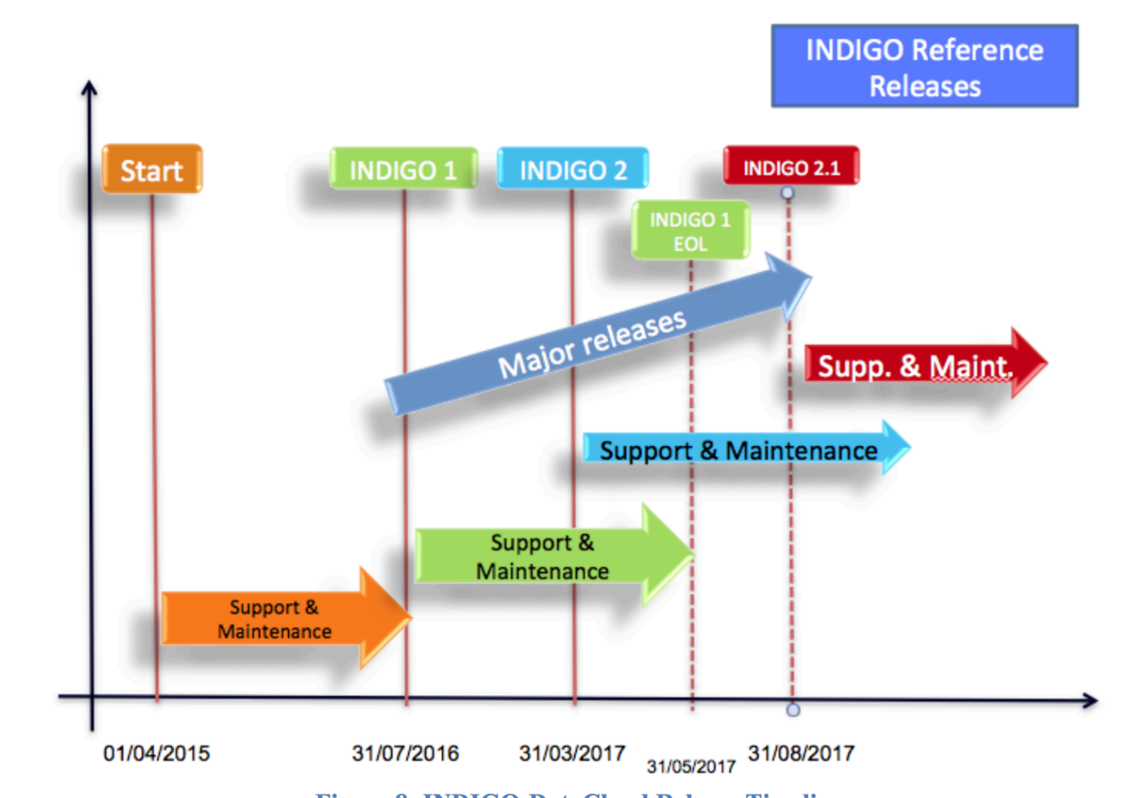


Figure 8: INDIGO-DataCloud Release Timeline

ElectricIndigo

- **NEW**: our second and final major software release, called **ElectricIndigo**
  - For technical details, see the parallel sections **on Thursday**
- **Fact sheet (https://www.indigo-datacloud.eu/service-component):**
  - 40 modular components, distributed via 170 software packages, 50 ready-to-use Docker containers
  - Operating systems: CentOS 7, Ubuntu 16.04
  - Cloud frameworks: OpenStack Newton, OpenNebula 5.x
  - Download it from the INDIGO-DataCloud Software Repository: http://repo.indigo-datacloud.eu/index.html

ElectricIndigo: Application-level Interfaces for Cloud Providers and Automated Service Composition

- Easily **port applications to public and private Clouds** using open programmable interfaces, user-level containers, and standards-based languages to **automate definition, composition and instantiation of complex set-ups**.

- **Typical questions**: How can I run my application on Cloud provider X? What if I want to use Docker but my provider does not support it? How do I automate the creation and management over public or private Clouds of dynamic clusters running multiple services?

ElectricIndigo: Flexible Identity and Access Management

- **Manage access and policies to distributed resources** using multiple methods such as **OpenID-Connect, SAML, X.509** digital certificates, through **programmable interfaces and web front-ends**.

- **Typical questions**: How can I manage access to distributed resources by users, identified through diverse methods? (e.g. Google ID, digital certificates) How should I modify / write my apps to benefit from that?

ElectricIndigo:
Data Management and Data Analytics Solutions

- **Distribute and access data** through multiple providers via **virtual file systems and automated replication and caching**, exploiting scalable, **high-performance data mining and analytics**.

- **Typical questions**: How can I automatically replicate datasets to multiple sites? Can I transparently access my distributed datasets from my app? Can I cache the most accessed data, so that it's close to where users need it? How do I instantiate clusters and databases for big data analysis?

ElectricIndigo:
Programmable Web Portals, Mobile Applications

- **Create and interface web portals or mobile apps,** exploiting **distributed data as well as compute resources** located in public and private Cloud infrastructures.

- **Typical questions**: How can I easily provide my app with a pluggable, extensible web front-end? Can this front-end interface with all the features provided by INDIGO? How can I write an INDIGO-enabled app for Android or iOS?

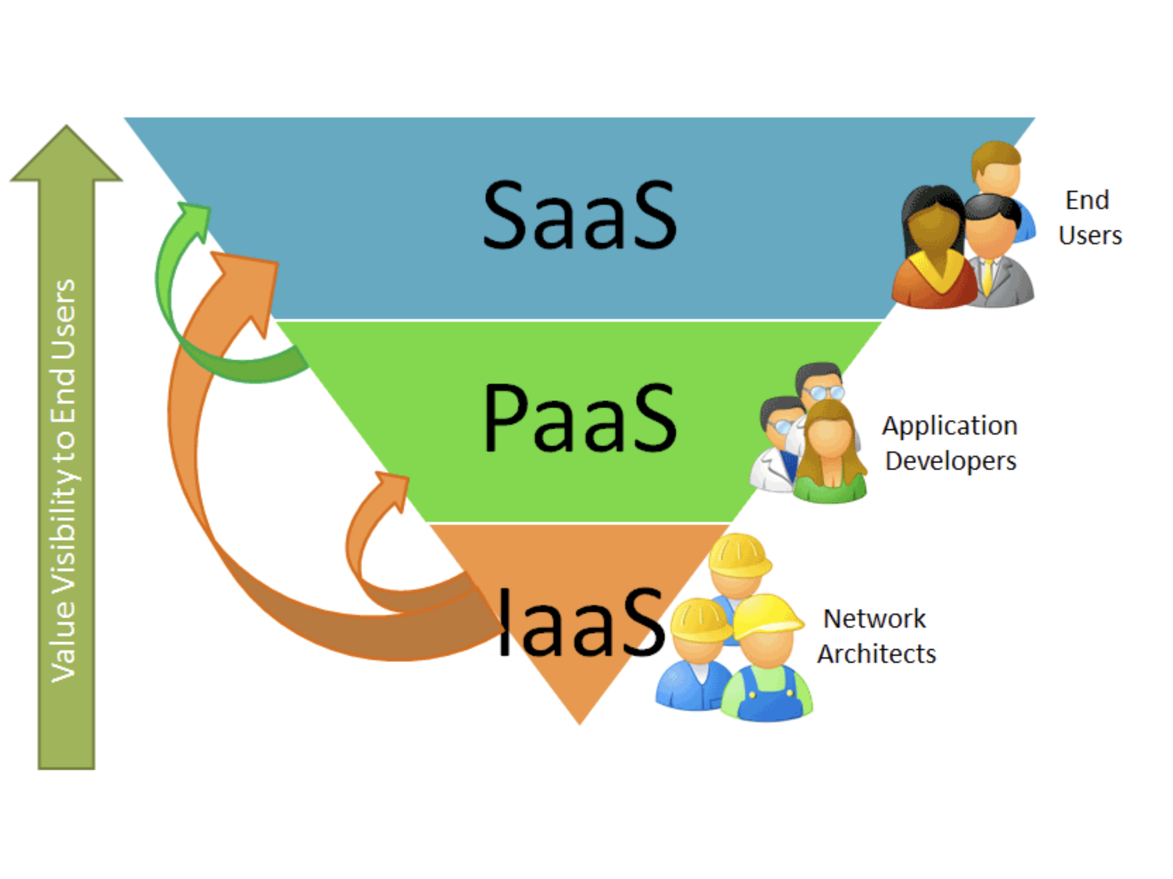
ElectricIndigo:
Enhanced and Scalable Services for Data Centers and Resource Providers

- **Increase the efficiency of existing Cloud infrastructures** based on OpenStack or OpenNebula through **advanced scheduling, flexible cloud / batch management, network orchestration** and interfacing of high-level Cloud services to existing storage systems.

- **Typical questions**: How can my cloud data centers provide flexible and fair scheduling policies for access to resources? How do I balance traditional vs. cloud resources in my data center? How do I connect novel INDIGO features to my existing systems? How can I manage storage Quality of Service?

How does this fit in a EOSC?

- We recognize that **value for users** (and hence, our main focus) is at the **upper layers**, not in the barebone e-infrastructural services.
  - But we also provide ways to optimize e-infrastructural services for resource providers
- So, we believe in more flexibility in choosing e-infra providers, resources and capabilities, **as long as**…
- … **users are empowered to easily express and implement their requirements through enabling services and components.**
- This is a movement that goes **well beyond the ”S” of Science** in the EOSC.



The role of INDIGO in a EOSC

- We see it in **three dimensions**:

1. **Support to scientific communities**: how can communities solve problems and come to results more effectively, more efficiently
2. **Support to innovation**: how can the EOSC profit from innovative solutions that were missing before INDIGO
3. **Support to evolution**: how can the INDIGO results and know-how be evolved in the future

INDIGO & EOSC: support to **communities**

- **Algae bloom modeling**
- **RNA sequencing with TRUFA**
- **Deploying an elastic cluster with INDIGO components**
- **Cloudified services for molecular dynamics**
- **A distributed archive system for the Cherenkov Telescope Array (CTA)**
- **The Large Binocular Telescope (LBT) distributed archive**
- **Ophidia for astronomical images calibration**
- **Launching POWERFIT and DISVIS VMs on the EGI FedCloud using INDIGO tools**
- **POWERFIT and DISVIS web portals: harnessing GPGPUs on the Grid using udocker**
- **Automated deployment of an Ophidia big data analytics cluster**
- **INDIGO at the Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information**
- **EGI and INDIGO integration**
- **ELIXIR-ITALY: developing a Galaxy instance provider platform**
- **Multidisciplinary Oceanic Information System**
- **Deploy Zenodo-based repository in the cloud using Marathon**
- **An on-demand analysis cluster for the CMS LHC experiment**

INDIGO & EOSC: support to innovation

- **Inter-site Networking with the INDIGO Virtual Router** – Demo booth, Tue morning
- ***bdocker* and *udocker*: two complementary approaches for the execution of containers in batch systems** – Demo booth, Tue afternoon
- **INDIGO-Datacloud meets the Open Telekom Cloud – a seamless and state-of-the-art hybrid cloud service for scientists** – Demo booth, Wed morning
- **The INDIGO Token Translation Service (WaTTS)** – Parallel session, Wed afternoon
- **Demo on the Token Translation client** – Parallel session, Wed afternoon
- **CDMI-based Storage Quality-of-Service Management** – Parallel session, Wed afternoon
- **Usage of the Cloud Fairshare Scheduler for OpenNebula** – Demo booth, Thu morning
- **Preemptible instances in the Cloud** – Demo booth, Thu morning
- **The INDIGO FutureGateway** – Demo booth, Thu afternoon
- **The orchestrator client** – Demo booth, Thu afternoon
- **ENES and Big Data Analytics: Ophidia + Kepler + Mobile Apps** – Demo booth, Thu afternoon

INDIGO & EOSC: support to **evolution**

- **How can INDIGO be sustained and evolved?**

1. Collaboration with commercial providers
2. Collaboration with other projects and initiatives
3. Open channel and forum
4. Submission of new projects

- **Join the Open Forum session** on Thursday afternoon, 14:30-16:00 to discuss details

New projects

- In the last round of the H2020 calls (March-April 2017), at least **5 proposals** were submitted that included key INDIGO components or their possible evolutions.
- Not all of these proposals may be approved, but it is interesting to note that **there is significant interest and request for solutions that originate from INDIGO**. If results are there, stakeholder engagement is strong, if ideas, requirements, architectures are valid, this interest will eventually find a way to be supported.

INDIGO & EOSC **in production**: >= TRL8

- For example, **INDIGO solutions and activities** are in the **EOSC-hub proposal** (a joint proposal between EGI, EUDAT and INDIGO-DataCloud)
- With **INDIGO components** such as Identity and Access Management, Token Translation, Virtual filesystems (Onedata), Advanced IaaS Services, the Infrastructure Manager, the INDIGO PaaS and its orchestrator, web front-end services, user-level containers
- And with **training, support, technical coordination, external liaison, stakeholder engagement, policy contributions**.

INDIGO & EOSC in evolution: < TRL8

- For example, **novel features** evolving INDIGO components are a key part of several proposals to the **EINFRA-21-2017 and ICT-16-2017 calls**:
  - Intelligent dataset distribution and data lifecycle management
  - Smart caching
  - Orchestrating Computing Workflows based on policy driven or adaptive data movements
  - Flexible metadata management for big data sets
  - Access to bare-metal resources on the Cloud
  - PaaS-Level access to HPC resources
  - Extensions to the INDIGO Orchestrator for hybrid IaaS deployments and scale out to 3rd party clouds
  - Extensions to the INDIGO Virtual Router Appliance
  - Real-time, streaming-based data ingestion and processing

INDIGO and External Projects: Components and Patches Merged in Upstream Open Source Projects

INDIGO - DataCloud

- OpenStack (https://www.openstack.org)
  - Nova Docker
  - Heat
  - OpenID-Connect for Keystone
  - Pre-emptible instances support (under discussion)
- OpenNebula (http://opennebula.org)
  - OneDock
- Infrastructure Manager (http://www.grycap.upv.es/im/index.php)
- Clues (http://www.grycap.upv.es/clues/eng/index.php)
- Onedata (https://onedata.org)
- TOSCA adaptor for JSAGA (http://software.in2p3.fr/jsaga/dev/)
- OCCI implementation for OpenStack (https://github.com/openstack/ooi)
- Extended AWS support for rOCCI in OpenNebula. Python and Java libraries for OCCI support.
- CDMI and QoS extensions for dCache (https://www.dcache.org)
- Workflow interface extensions for Ophidia (http://ophidia.cmcc.it)
- OpenID Connect Java implementation for dCache (https://www.dcache.org)
- MitreID (https://mitreid.org/) and OpenID Connect (http://openid.net/connect/) libraries

More this week

- At the **plenaries on Thursday morning**, we discuss the societal impact of the EOSC, exploitation experiences of INDIGO solutions in open source initiatives, big research communities and industry.
- **On Thursday afternoon**, we debate how INDIGO services can be part of channels, open forums, complementing services offered by e-infrastructures, research infrastructures and private cloud providers, and we delve into the technical details of the ElectricIndigo release.
- **On Friday**, we elaborate on data ingestion implemented with INDIGO tools, examine INDIGO solutions at the IaaS, PaaS and SaaS levels, and discuss new ideas and initiatives to extend INDIGO components.
- **Take the time to explore what's on show at the Summit, talk to people, provide input, ask questions… and enjoy beautiful Catania!**

Conclusions

- In 24 months, the INDIGO-DataCloud project has realized a **comprehensive involvement of many Research Communities and providers** for the definition and tracking of requirements.
- We identified **technology gaps** linked to several concrete use cases, defined, published and implemented the **overall INDIGO architecture**.
- After early demonstrations and beta software previews, we **produced two major software versions and 9 minor updates**, releasing 40 open modular components. We did that exploiting key European know-how, reusing and extending open source software, and contributing to upstream projects. We established software development and management processes, and defined development and pre-production distributed testbeds.
- **Production deployment of many applications making use of the INDIGO software** is well underway, and INDIGO components have been proposed for production use in big infrastructures, commercial companies, external projects.
- **Several opportunities for further exploitation of INDIGO components** are being explored and implemented, in the context of the EOSC and beyond.

INDIGO - DataCloud

Thank you

https://www.indigo-datacloud.eu

Better Software for Better Science.

@indigodatacloud

www.indigo-datacloud.eu

https://www.facebook.com/indigodatacloud/