CTS2014 Tutorial:

Cloud Based Federated Infrastructure for Big Data e-Science and Collaboration

(European focus and examples)

Yuri Demchenko
System and Network Engineering, University of Amsterdam

CTS2014 Conference
19-23 May 2014, Minneapolis, USA

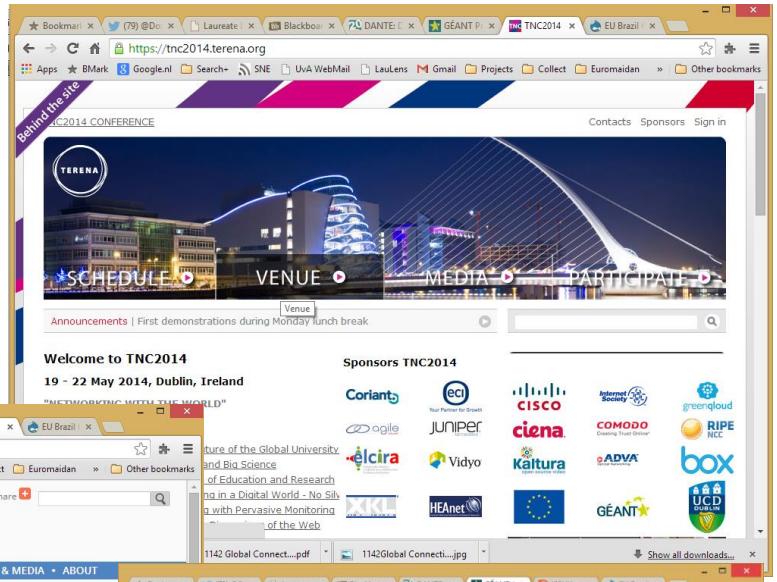
Outline

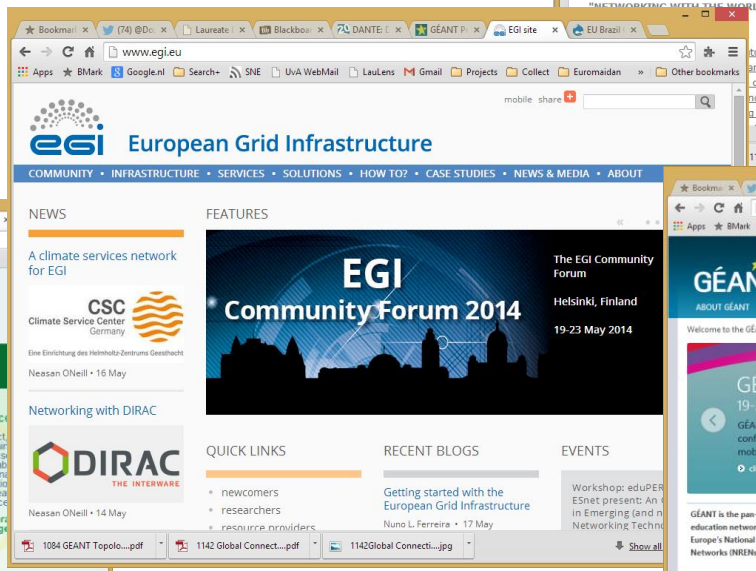
- e-Science and Big Data challenges
  - The 4th Paradigm, Big Data and long-tale science
  - European Research Areas (ERA) and projects
    - Collaboration and information sharing
- e-Science and Research Infrastructures as a basis for wide collaboration in science
  - EU-Brazil Cloud Connect Project and use cases
  - European Grid Infrastructure: EGI Federated Cloud Infrastructure
  - GEANT European Research and Education Network
- Scientific Data Infrastructure for Big Data
- Federated security models in cloud
  - Legacy Virtual Organisations (VO) based federated access control infrastructure
  - Generic Federated Access Control and Identity Management in cloud
- Implementation in the GEANT Infrastructure
- Discussion

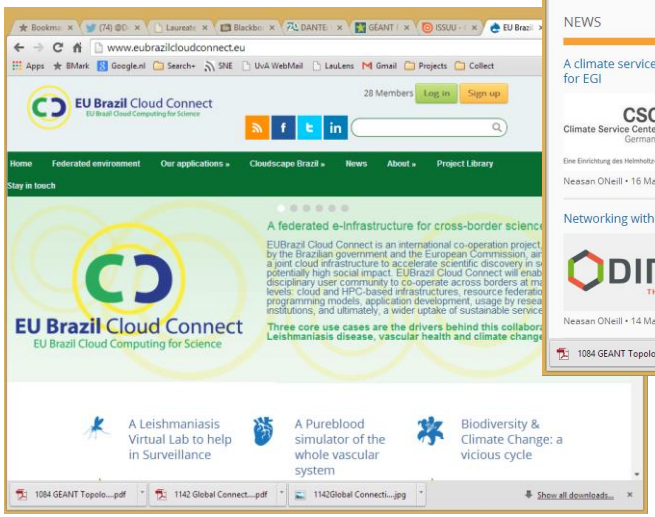
http://www.uazone.org/demch/presentations/cts2014tutorial02.pdf

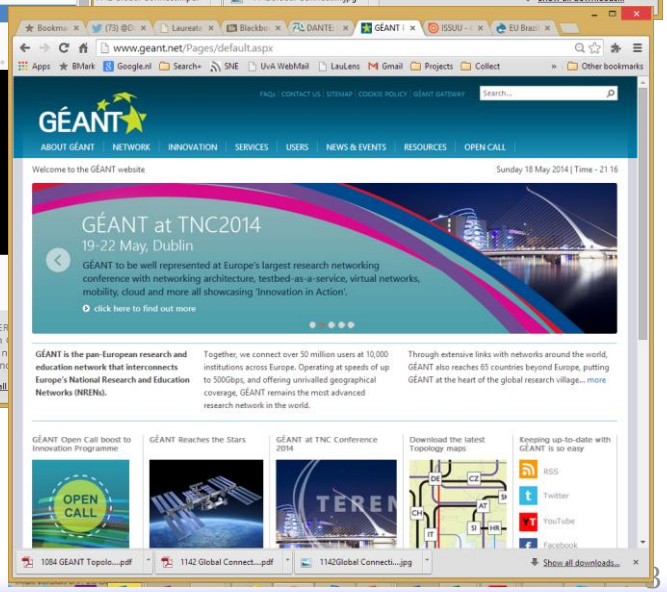
This week 19-23 May 2014: Conferences and Events

- TERENA Networking Conference TNC2014
  - https://tnc2014.terena.org/
- European Grid Infrastructure (EGI)
  - http://cf2014.egi.eu/
- EU-Brazil Cloud Connect Project
  - http://www.eubrazilcloudconnect.eu/
- GEANT Network for Research and Education in Europe
  - http://www.geant.net/MediaCentreEvents/Events/GEANT_at_TNC_2014/Pages/Home.aspx





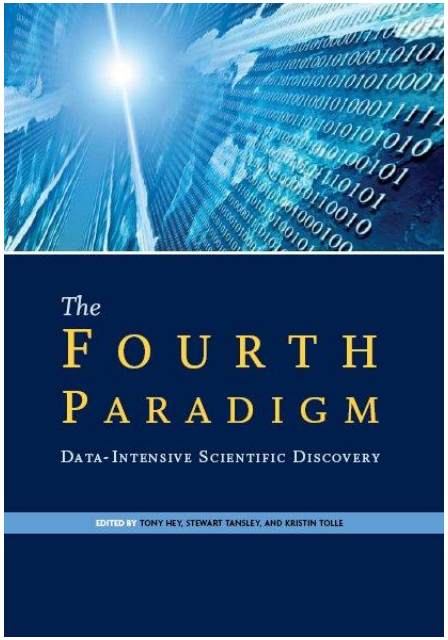




Yuri Demchenko – Professional Summary

- Graduated from National Technical University of Ukraine “Kiev Polytechnic Institute” (KPI) in Instrumentation and Measurement (aka Industry Automation)
  - Candidate of Science (Tech) – Dissertation on System Oriented Precision Generators (1989)
- Teaching at KPI 1989-1998 – Computer Networking, Internet Technologies, Security
- Professional work in Internet technologies since 1993
- Work at TERENA (Trans-European R&E Networking Association) – 1998-2002
- Work at UvA with SNE group – since 2003
  - Main research areas: Cloud Computing, Big Data Infrastructures, Application and Infrastructure Security, Generic AAA&Authorisation, Grid and collaborative systems
  - EU Projects: GEYSERS, GEANT3, Phosphorus, EGEE I-II, Collaboratory.nl
  - Standardisation activity – IETF, Open Grid Forum (OGF) – ISOD-RG chairing, NIST Cloud Collaboration, NIST Big Data WG, ISO/IEC Big Data Study Group
  - Now/2014: Big Data Architecture, Big Data Security, Big Data Curriculum development

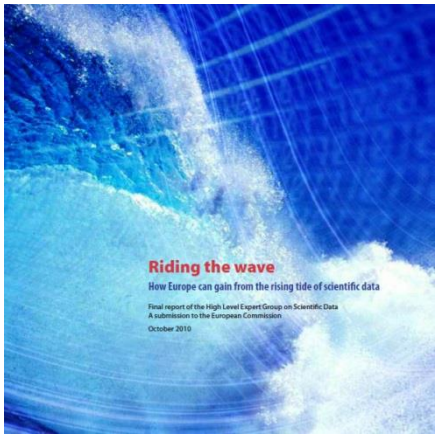
e-Science and Big Data: Seminal works, High level reports, Initiatives



The Fourth Paradigm: Data-Intensive Scientific Discovery.

By Jim Gray, Microsoft, 2009. Edited by Tony Hey, et al.

http://research.microsoft.com/en-us/collaboration/fourthparadigm/



Riding the wave: How Europe can gain from the rising tide of scientific data.

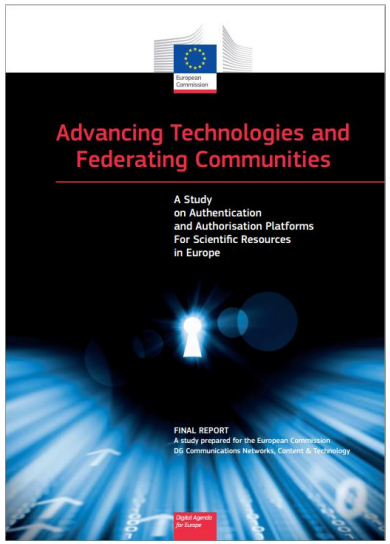
Final report of the High Level Expert Group on Scientific Data. October 2010.

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

RDA RESEARCH DATA ALLIANCE

Research Data Sharing without barriers

https://www.rd-alliance.org/



AAA Study: Study on AAA Platforms For Scientific data/information Resources in Europe, TERENA, UvA, LIBER, UinvDeb.

NIST Big Data Working Group (NBD-WG)
https://www.rd-alliance.org/

The Fourth Paradigm of Scientific Research

1. Theory, hypothesis and logical reasoning
2. Observation or Experiment
   - E.g. Newton observed apples falling to design his theory of mechanics
   - But Gallileo Galilei made experiments with falling objects from the Pisa leaning tower
3. Simulation of theory or model
   - Digital simulation can prove theory or model
4. Data-driven Scientific Discovery (aka Data Science)
   - More data beat hypnotized theory
   - e-Science as computing and Information Technologies empowered science

Big Data and Data Intensive Science - The next/current technology focus

- Based on e-Science concept and entire information and artifacts digitising
  - Requires also *new information and semantic models* for information structuring and presentation
  - Requires new research methods using large data sets and data mining
    - Methods to evolve and results to be improved
- Changes the way how the modern research is done (in e-Science)
  - Secondary research, data re-focusing, linking data and publications
- Big Data requires ***a new infrastructure*** to support both distributed data (collection, storage, processing) and metadata/discovery services
  - High performance network and computing, distributed storage and access
  - Cloud Computing as a native platform for distributed dynamic virtualised (data supporting) infrastructure
  - Demand for trusted/trustworthy infrastructure

e-Science Features

- ***Automation*** of all e-Science processes including data collection, storing, classification, indexing and other components of the general data curation and provenance
- ***Transformation*** of all processes, events and products ***into digital form*** by means of multi-dimensional multi-faceted measurements, monitoring and control; digitising existing artifacts and other content
- Possibility to ***re-use the initial and published research data*** with possible data re-purposing for secondary research
- ***Global data availability*** and access over the network for cooperative group of researchers, including wide public access to scientific data
- Existence of necessary infrastructure components and management tools that allows fast ***infrastructures and services composition, adaptation and provisioning on demand*** for specific research projects and tasks
- ***Advanced security and access control*** technologies that ensure secure operation of the complex research infrastructures and scientific instruments and allow creating ***trusted secure environment*** for cooperating groups and individual researchers.

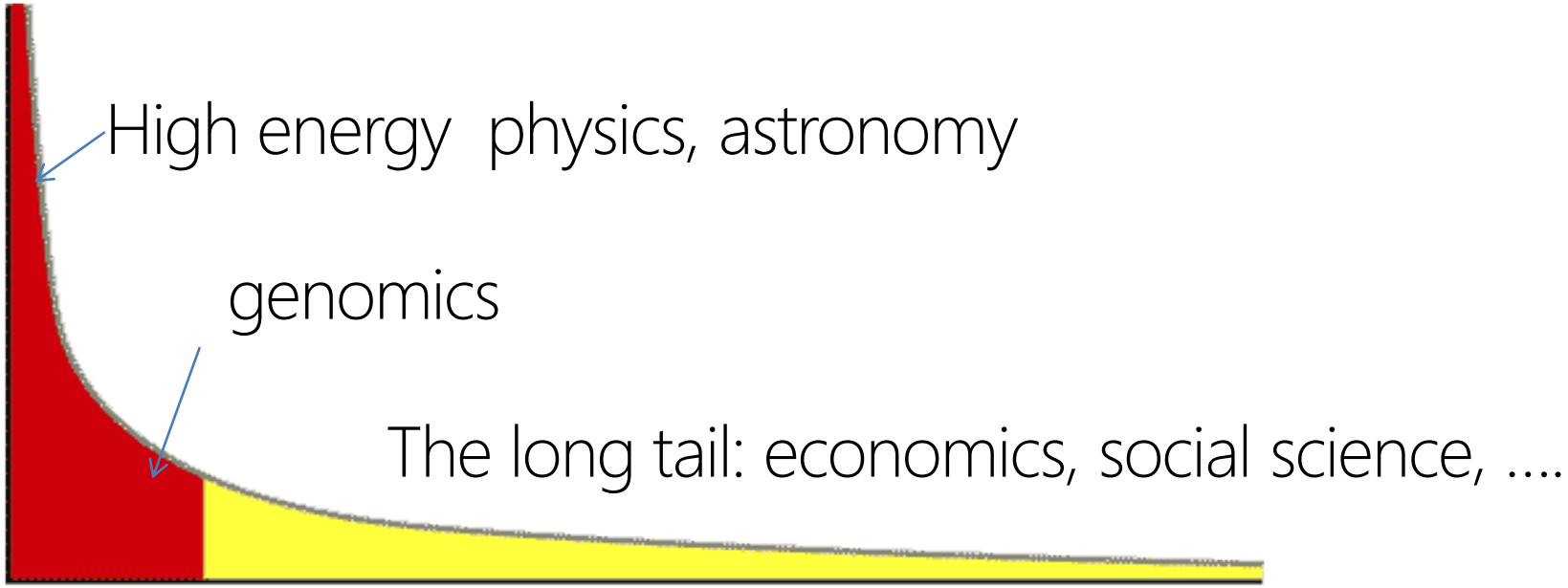
Modern e-Science in search for new knowledge as a Big Data technology driver

Scientific experiments and tools are becoming bigger and heavily based on data processing and mining

- 3 V of Big Data challenges for Scientific Data Infrastructure (SDI)

- Volume – Terabyte records, transactions, tables, files.
  - LHC – 5 PB a month (now is under re-construction)
  - LOFAR, SKA – 5 PB every hour, requires processing asap to discard non-informative data
  - Large Synoptic Survey Telescope (LSST) - 10 Petabytes per year
  - Genomic research – x10 TB per individual
  - Earth, climate and weather data
- Velocity – batch, near-time, real-time, streams.
  - LHC ATLAS detector generates about 1 Petabyte raw data per second, during the collision time about 1 ms
- Variety – structures, unstructured, semi-structured, and all the above in a mix
  - Biodiversity, Biological and medical, facial research
  - Human, psychology and behavior research
  - History, archeology and artifacts

The Long Tail of Science (aka “Dark Data”)



- Collectively “Long Tail” science is generating a lot of data
  - Estimated as over 1PB per year and it is growing fast with the new technology proliferation
- 80-20 rule: 20% users generate 80% data but not necessarily 80% knowledge

Source: Dennis Gannon (Microsoft)
NIST Big Data Workshop, 2012

European Research Area (ERA) - Coordination

- European Commission – *but not only*
  - *Horizon2020 new EU Framework Program 2014-2020 to support Research and Innovation in Research and Industry*
- EIROforum – European Intergovernmental Research Organisation
  - Profile committees organised by scientific domain
- ESFRI – European Strategy Forum for Research Infrastructure
  - Coordinates projects and funding for Research Infrastructures (RI)
- eIRG – e-Infrastructure Reflection Group
  - High level policy development for Europe on e-Infrastructure
- EEF - European e-Infrastructure Forum
  - Principles and practices to create synergies for distributed Infrastructures
- TERENA and DANTE
  - GEANT high performance European Research and Education Network
  - REFEDS – Research and Education Federations
- LIBER – Association of European libraries
  - Growing role of scientific libraries including access to research information
- Research Data Alliance (RDA)
  - Joint initiative by ERA/EC, NSF, NIST

Big Data Science and European Research Areas (1)

- High Energy Physics (HEP)
  - Running experiment on LHC and infrastructure WLCG (Worldwide LHC Grid)
    - Already producing PBytes of information
    - Worldwide distribution and processing
  - CERN and national HEP centers
- Low Energy Physics and Material Science (photon, proton, laser, spectrometry)
  - Number of research facilities serving international communities
  - Multiple short projects producing TBytes of information
    - Experimental data storage, identification, trusted access to multiple users (including public and private researchers)
- Earth, weather and space observation
  - Climate research and Earth observation
    - With new 4? satellites to be launched starting 2017 to produce PBytes monthly
  - ESA (European Space Agency)

Big Data Science and European Research Areas (2)

- Life science and biodiversity (Genomic, Biomedical and Healthcare research)
  - Human genome (EMBL-EBI)
    - Currently centralised databases but evolving to distributed
    - ELSI data - Special requirements to data integrity and privacy
  - Living species and biodiversity
    - Mobile/field access, filtering and on-demand computing
    - Public contribution, vocational or citizen researchers
  - Numerous local/offline databases to be brought online
  - Projects: ENVRI, LifeWatch, ELIXIR, HelixNebula
- Humanities (History, languages, human behaviour)
  - Rediscovering research with total information digitising
    - Expected huge amount of data to digitise all human heritage
  - Very spread research community
  - Projects: CLARIN, DARIAH, EUDAT
- Outreach and cooperation with developing research communities
  - Brazil, China, Africa

Existing and emerging Europe wide SDI

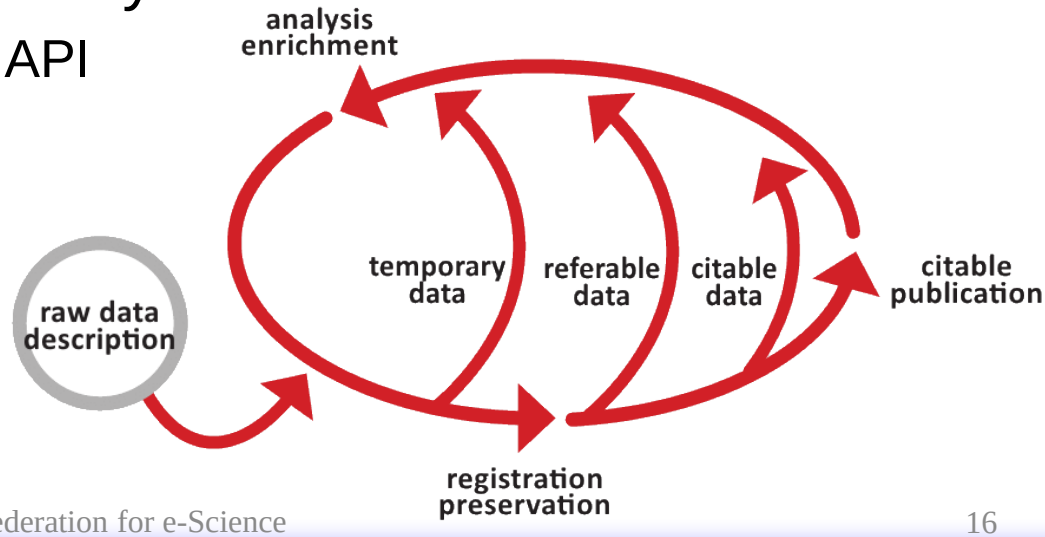
- WLCG – Worldwide LHC Grid (CERN, Geneva)
- EGI – European Grid Infrastructure (successor of the EGEE project)
  - Operational Grid infrastructure serving around 10,000 researches worldwide
  - Published “Seeking new horizons: EGI’s role for 2020”
  - ***Federated Cloud Infrastructure*** provides an infrastructure platform for operational and legacy Grid services
- PRACE – Partnership for Advanced Computing in Europe
- HELIX Nebula – The Science Cloud (prospective cloud based SDI for ERA)
  - Private Partnership Project with wide industry participation (limited EC/FP7 support)
- Growing Research Infrastructures for different research communities
  - CLARIN, EUDAT, LifeWatch, ELIXIR, etc.
    - Less technology and more subject focused

Open Access to Scientific Publications

- EC initiative on Open Access scientific publications from publicly funded projects
  - Included into Declaration from the H2020 Rome meeting (2012)
  - Approx 3500 publicly funded ROs and 2000 privately funded ROs
  - Special funding scheme for reimbursing publications
  - Issues with China, India, Russia compliance to OA principles
    - Consultation at high governmental level
- OpenAIRE project exploring models for open access to publications
  - PID (Persistent ID for data), ORCHID (Open Researcher ID), Linked data
- Community initiative - Panton Principles for Open Data in Science (http://pantonprinciples.org/)

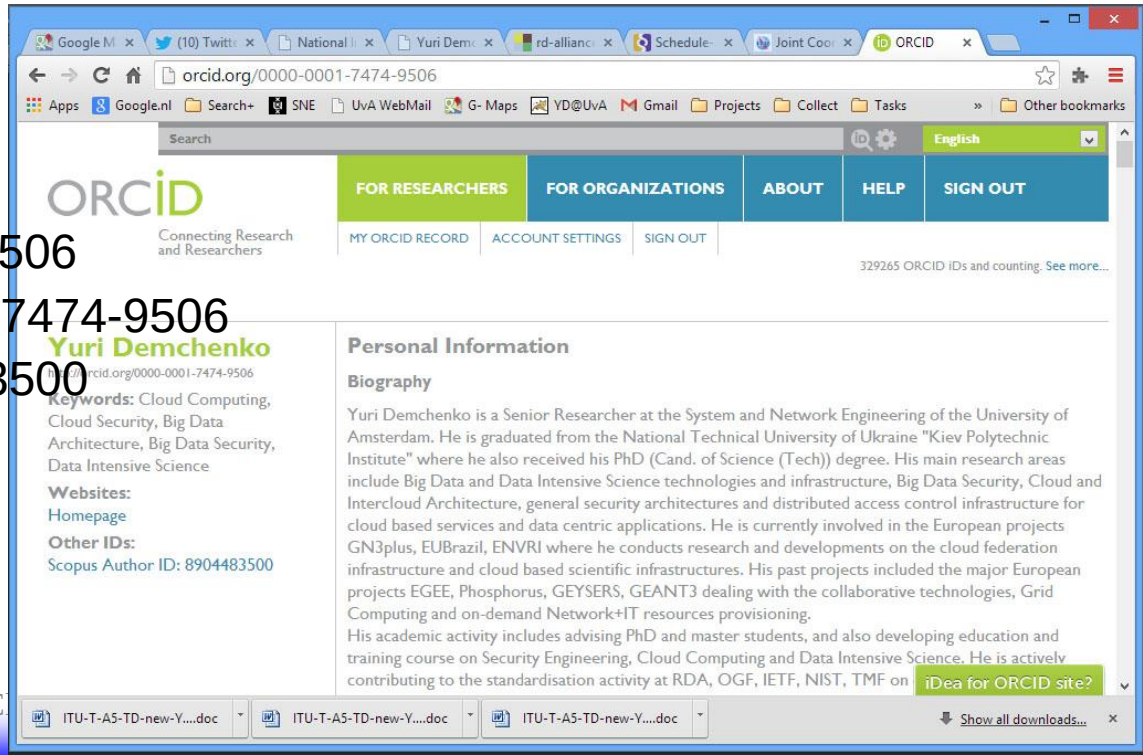
Persistent Identifier (PID)

- PID – Persistent Identifier for Digital Objects
  - Managed by European PID Consortium (EPIC) http://www.pidconsortium.eu/
  - Superset of DOI - Digital Object Identifier (http://www.doi.org/)
  - Handle System by CNRI (Corporation for National Research Initiatives) for resolving DOI (http://www.handle.net/)
- PID provides a mechanism to link data during the whole research data transformation cycle
  - EPIC RESTful Web Service API published May 2013

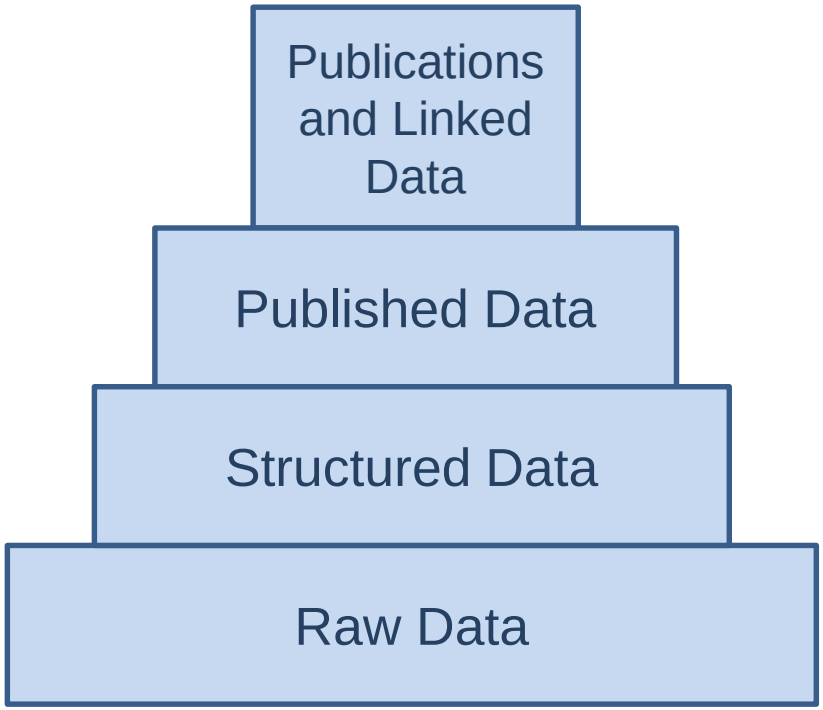


ORCID (Open Researcher and Contributor ID)

- ORCID is a nonproprietary alphanumeric code to uniquely identify scientific and other academic authors
  - Launched October 2012
- ORCID Statistics – May 2014
  - Live ORCID IDs 511, 203 (October 2013 - 329,265)
  - ORCID IDs with at least one work 121,529 (October 2013 - 79,332)
  - Works 2,205,971
  - Works with unique DOIs 1,267,083
- Personal ORCID
  - ORCID 0000-0001-7474-9506
  - http://orcid.org/0000-0001-7474-9506
  - Scopus Author ID 8904483500

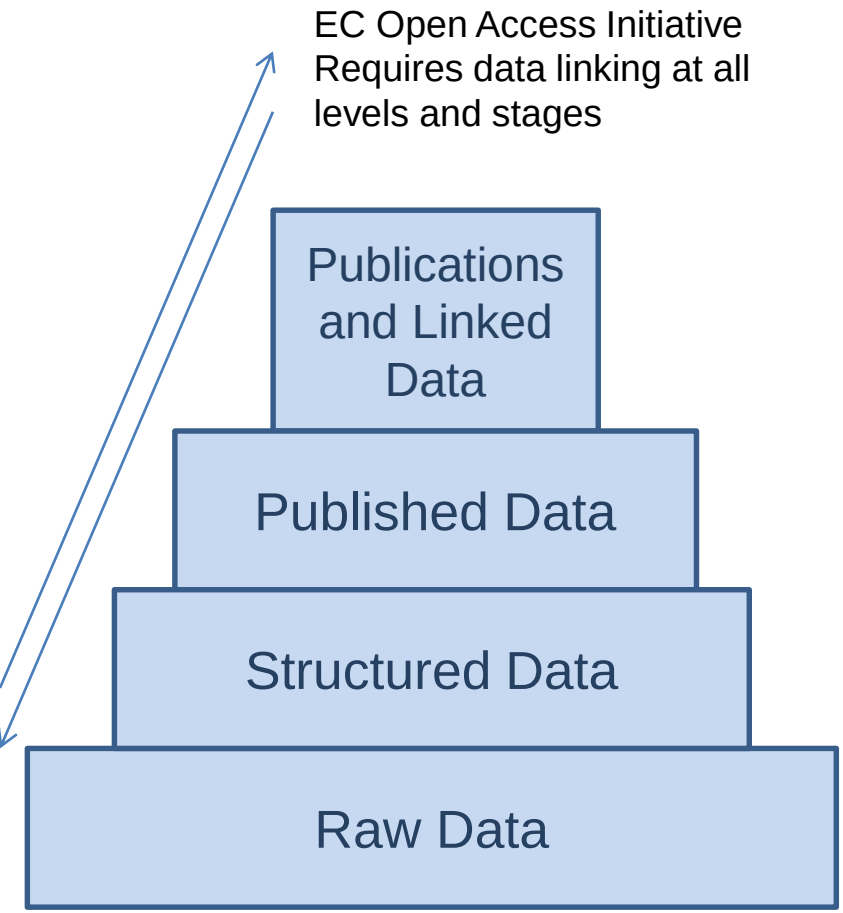


Scientific Data Types



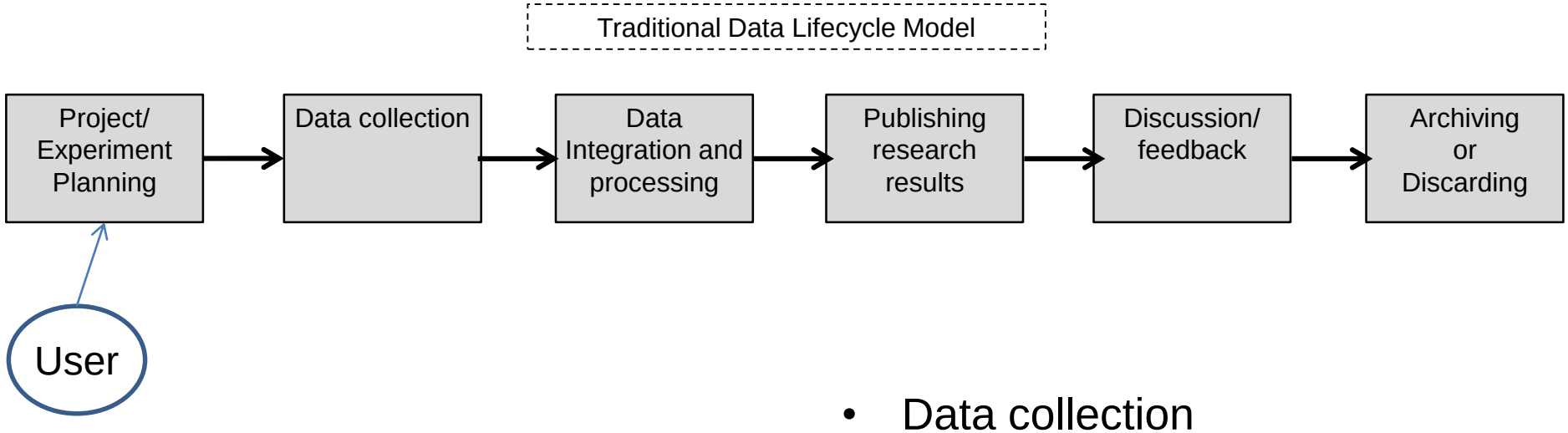
- **Raw data** collected from observation and from experiment (according to an initial research model)
- **Structured data** and datasets that went through data filtering and processing (supporting some particular formal model)
- **Published data** that supports one or another scientific hypothesis, research result or statement
- **Data linked to publications** to support the wide research consolidation, integration, and openness.

Scientific Data Types



- **Raw data** collected from observation and from experiment (according to an initial research model)
- **Structured data** and datasets that went through data filtering and processing (supporting some particular formal model)
- **Published data** that supports one or another scientific hypothesis, research result or statement
- **Data linked to publications** to support the wide research consolidation, integration, and openness.

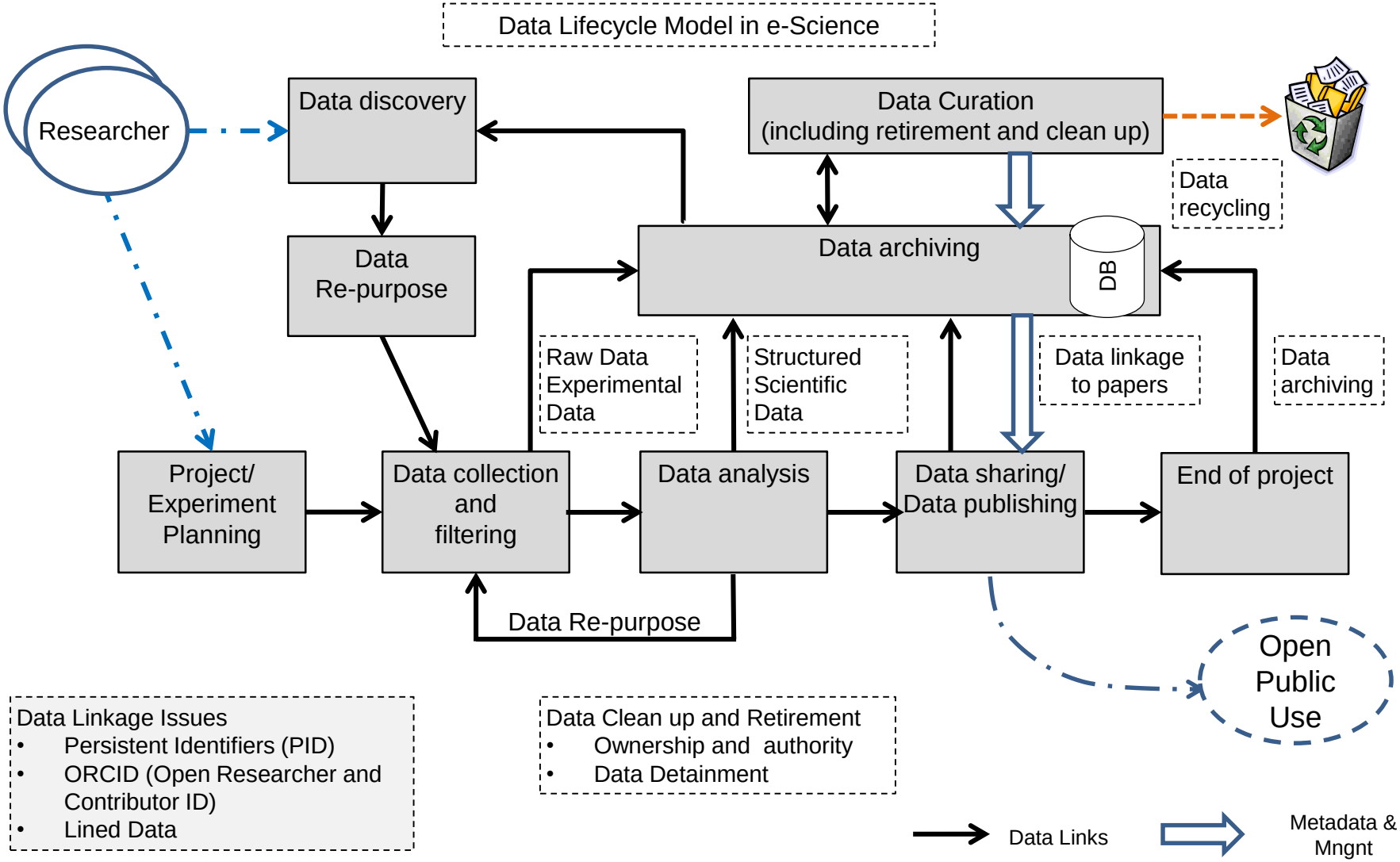
Traditional Data Lifecycle Model - I



- Data collection
- Data processing
- Publishing research results
- Discussion
- Data and publications archiving

Lack of initial data preservation and data linking to publications

Data Lifecycle Model in e-Science – II



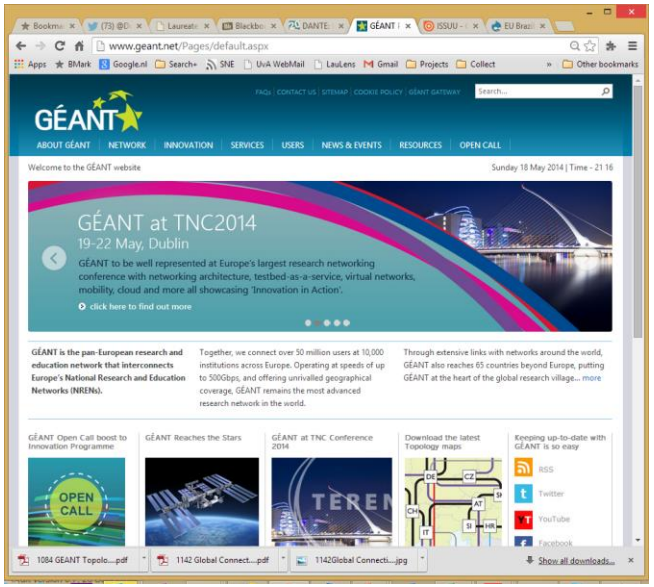
European Research Infrastructure: Examples and Projects

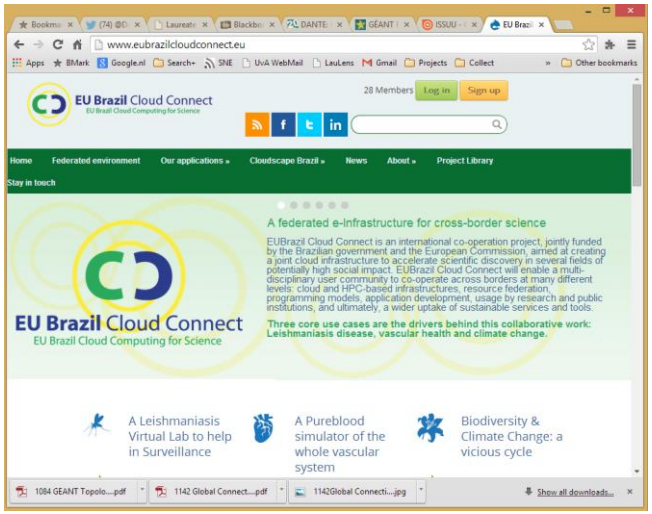
Scientific Applications

Cloud/Grid Infrastructure

Network Infrastructure

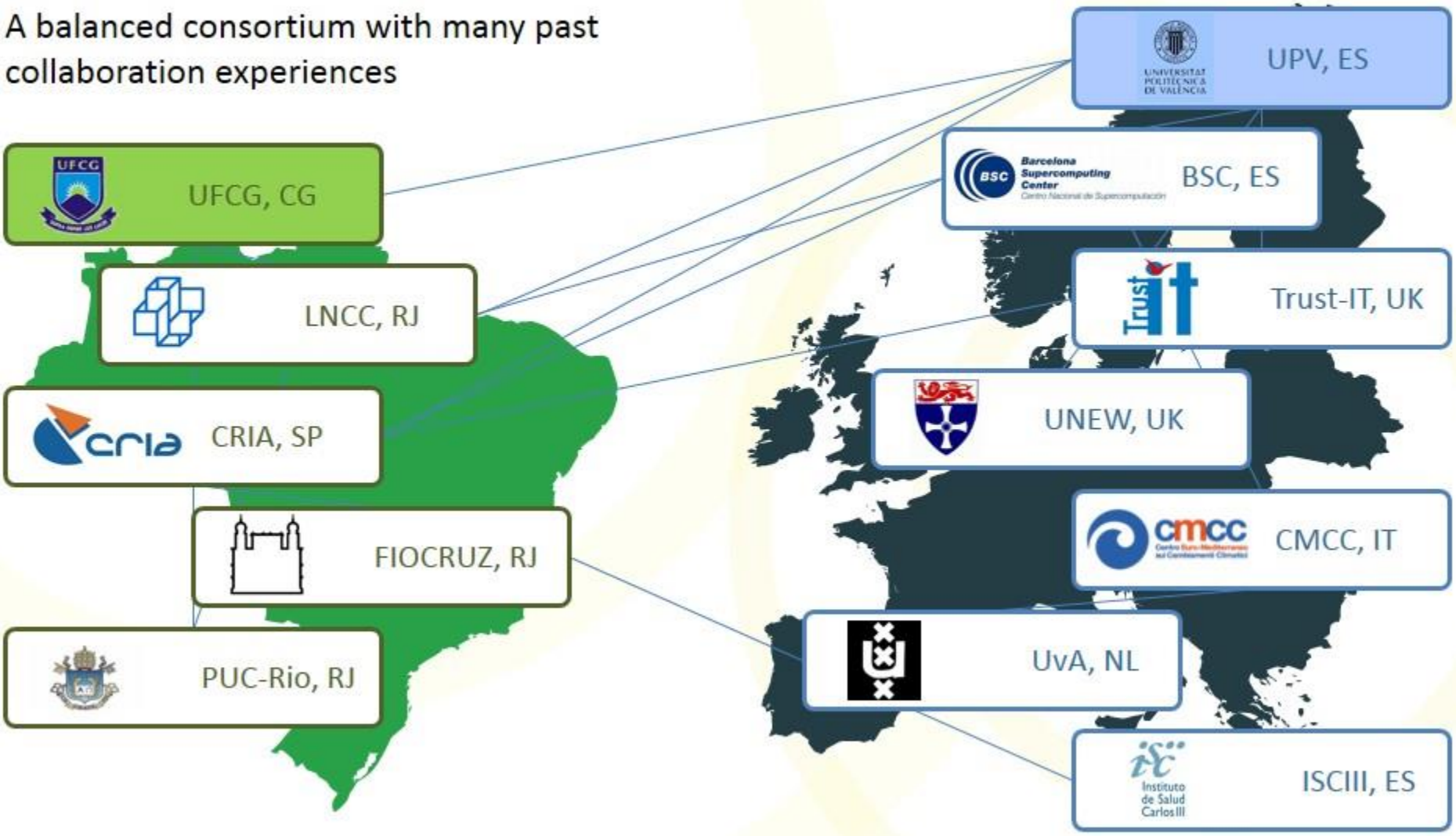
- EU-Brazil Cloud Connect Project
  - http://www.eubrazilcloudconnect.eu/
- European Grid Infrastructure (EGI)
  - http://www.egi.eu/
- GEANT Network for Research and Education in Europe
  - http://www.geant.net/





EUBrazilCC consortium: Years of collaboration

A balanced consortium with many past collaboration experiences



EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

EU Brazil Cloud Connect

Aims and benefits

- The main objective is the creation of a federated e-infrastructure for research using a user-centric approach.
- To achieve this, we need to pursue three objectives:
  - **Adaptation** of existing applications to tackle **new scenarios** emerging from cooperation between Europe and Brazil relevant to both regions.
  - Integration of frameworks and programming models for **scientific gateways and complex workflows**.
  - Federation of resources, to build up **a general-purpose infrastructure comprising existing and heterogeneous resources**
- Additionally, EUBrazilCC will: perform an active **dissemination** campaign, analyse **innovation**, foster the involvement of Brazilian institutions in **cloud standards definition**, and bring the EU Cloudscape series to broader international audience.

Slides courtesy of EUBrazilCC Project

EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

A Complete International Infrastructure

- Leveraging from existing Computing and Storage Resources
  - A total of >5500 CPU cores and >500TB and MareNostrum for UC2.
- The integration of different frameworks for cloud resources and services federation

FogBow

- FogBow is a MW for opportunistic usage of underused resources, evolved from OurGrid.
- FogBow follows a bartering model to provide access to cloud federations

jitclouds.lsd.ufcg.edu.br

CSGRID

- Management of distributed computational resources
- Executes of different versions of applications in distributed & heterog. environments.

jira.tecgraf.puc-rio.br/confluence

COMPSs

- Programming framework for the execution of parallel applications on distributed infrastructures.
- It discovers the parallelism through the dependencies among tasks, constructed from function calls.

http://compss.bsc.es/

Slides courtesy of EUBrazilCC Project

EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

… and Programming Platform

- Providing services to integrate pipelines with multiple depending components and Big Data analysis

Parallel Data Analytics (PDAS)

- big data analytics framework solutions to manage large volumes of data, perform time series analysis, data aggregation, transformation, etc.
- A hierarchical storage architecture to manage multidimensional data for scientific domains.

(presentation at EGI-CFG2014 scheduled on May 21st, 4pm, Room 10, *Session New data management solutions for EGI*)

eScienceCentral

- A workflow-based platform for data analysis.
- It supports applications coded into blocks in "R", java, octave or javascript.
- Supports public and on-premises clouds.

www.esciencecentral.co.uk

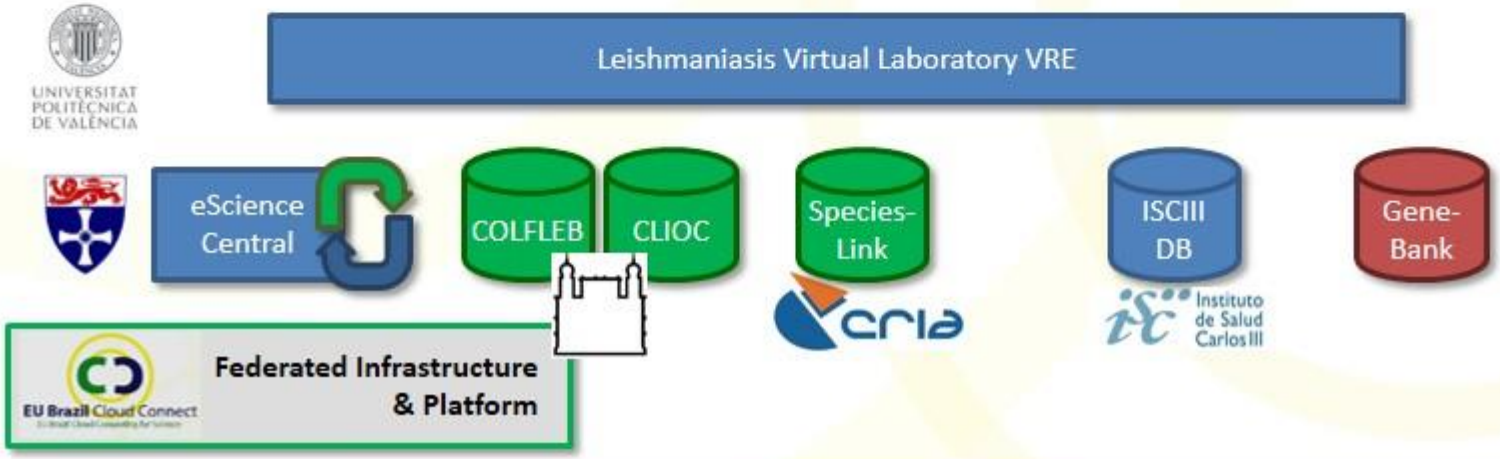
My Sci. Cloud (mc2)

- Platform for the development and provisioning of scientific applications running on cloud infrastructures.
- Environment for scientific gateways.

www.lncc.br/sinapad/projectmanager/public/projects/gt-mcc

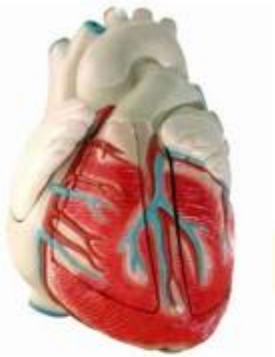
Use Case 1: Leishmaniasis Virtual Laboratory

- **Led by** ISCIII / FIOCRUZ.
- **Objective**: Improve knowledge on the distribution and susceptibility of epidemiology outburst in Leishmaniasis Disease
- **Technical Challenge**: Easy access to computing and data federation for applications defined as workflows.
- **International Added Value**: Linking data from Brazilian and European leaders and complementary databases and develop a Virtual Research Environment for integrating workflows for epidemiology risk modelling.

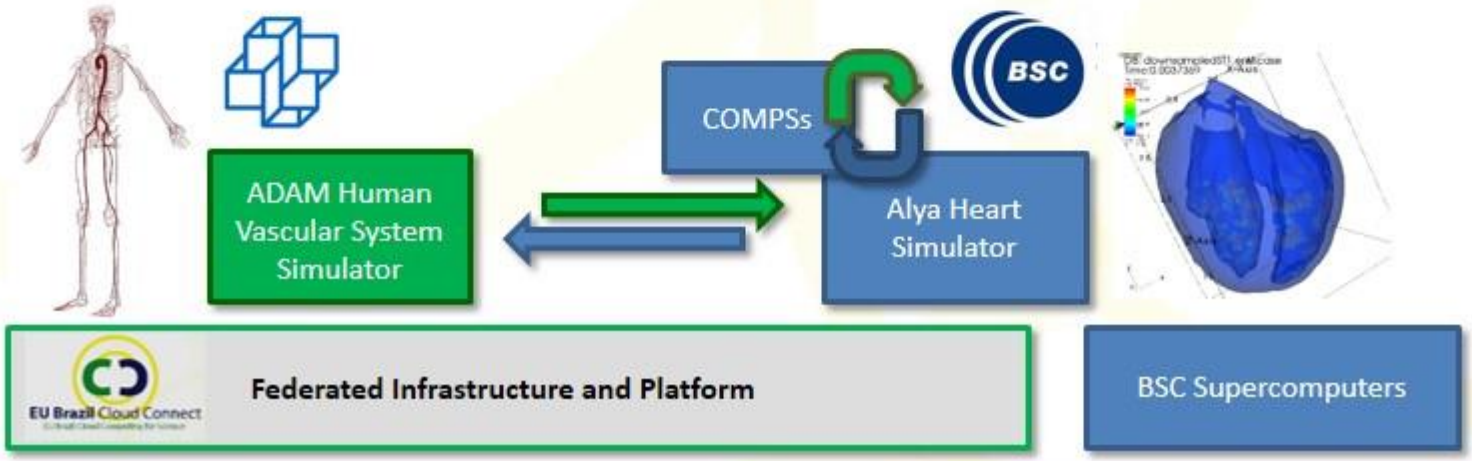


EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

Use Case 2: Heart Simulation



- **Led by:** BSC & LNCC.
- **Objective**: Increase the accuracy of blood simulation.
- **Technical Challenge**: Integrate Supercomputing and Cloud computing applications.
- **International Added Value**: Linking boundary conditions of the ADAM Vascular system to the ALYA multilevel heart simulator to achieve beyond the state-of-the-art simulation of the whole Human Vascular System Simulation.



EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

Use Case 3: Biodiversity and Climate Change

- **Led by**: CMCC & UFCG.
- **Objective**: Understand the impact of climate change on terrestrial biodiversity through two workflows based on Earth observation and ground level data.
- **Technical Challenge**: Integrate parallel data analysis with other processing workflows in a geographically distributed environment.
- **International Added Value**: Integration of biodiversity data and modelling with multispectral and remote sensing data for studying the cross-correlation of biodiversity and climate change.

Slides courtesy of EUBrazilCC Project

Strongly Related to other EU and BR projects

- HelixNebula (helix-nebula.eu)

  EUBrazilCC is already accepted as one interoperability testing use case.

- EGI-InsPIRE (www.egi.eu)

  BSC already contributing to the EGI Federated cloud Task Force (COMPSs+PMES).

- LifeWatch (www.lifewatch.eu)

  UvA and CSIC are the leaders of this ESFRI, and have interest in both UC1 and especially in UC3.

- CloudWATCH (www.cloudwatchHUB.eu)

  Definition of a cloud standard profile. An opportunity to use the innovation platform, run by DIGITALEUROPE to showcase Brazilian developments In cloud computing.

- SERPRO/Dataprev/Telebras Government Cloud Initiative

  LNCC leads this project to develop pilots on cloud computing for government IT companies.

- INCT-MACC - National Institute for Science and Technology in Medicine Assisted by Scientific Computing

  LNCC is the leader of this project.

- INCT- HVFF Brazilian Virtual Herbarium, INCT Negleted Diseases, and the Brazilian System for Biodiversity Information

  Participation of CRIA and LNCC

EGI Federated Cloud

EGI – European Grid Initiative

- Follow up after EGEE project (2004-2010) to create a Grid infrastructure to support LHC experiment in CERN
  - Worldwide LHC Grid (WLCG) http://wlcg.web.cern.ch/
- Legacy federated resources sharing and security around VO (Virtual Organisations)
- Currently moving Grid applications to Cloud platform

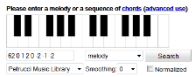
EGI Participation – Feb 2014

Cyfronet | FZJ | OeRC | EGI.eu | CESNET | GWDG | BIFI | KISTI | INFN-BARI

CNRS | SAGrid | KTH | FCTSG | CETA | IGI | RADICAL | STFC | BSC | ISRGrid

IN2P3 | Masaryk | INFN-CNAF | CESGA | SARA | IFCA | SZTAKI | GRNET | DANTE | SRCE

LMU | IPHC | IISAS | SixSq | 100%IT | CSC | IFAE | DESY



esa

wenmr

SCI-BUS

BioVeL

ENVRI

Calcium Science Gateway Framework

CLARIN
Common Language Resources and Technology Infrastructure

LifeWatch

Gaia

Members

- 142 individuals
- ~37 institutions
- 20 countries (EU & non-EU)

Stakeholders

- 23 Resource Providers
  - 13 production
  - 10 Certified
- 10 Technology Providers
- 10 User Communities
- 4 Liaisons

Technologies

- OpenNebula
- StratusLab*
- OpenStack
- Synnefo
- Cloudstack
- PERUN
- SlipStream
- APEL
- GOCDB

EGI Mission and Principles

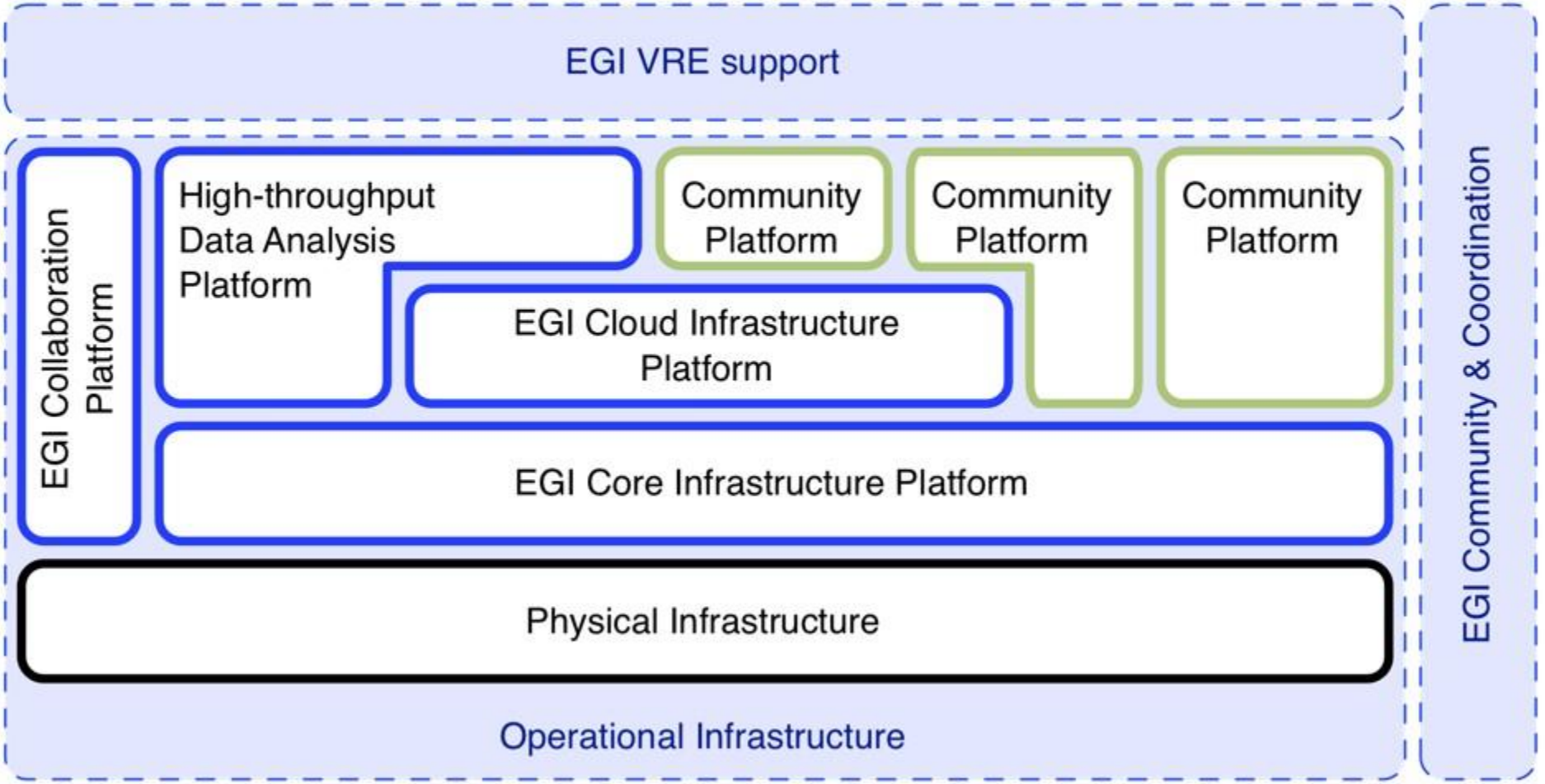
MISSION: To support international researcher collaborations from all disciplines with the reliable and innovative ICT services they need to accelerate science excellence

- Natural and physical sciences
- Medical and health sciences
- Engineering and technology

EC EGI-InSPIRE project (2010-2014) http://www.egi.eu/case-studies/

- Uniform access to heterogeneous data and compute services
  - Grid and Cloud platforms
- Federation of services from
  - Publicly funded infrastructures
  - Institutional infrastructures
  - Commercial providers (incl. partnership with HelixNebula)
    - Free at point of delivery/pay per use

EGI Strategy & Technical Architecture



EGI Services for long tail and big science

Individual Researchers & Teams

Research Communities & Institutions

Resource Centres & Service providers

High-Throughput Data Analysis

For management and analysis of large datasets and execution of thousands of computational tasks

European federation of publicly-funded clusters

- **Grid Compute**: computational jobs
- **Grid Storage**: store/access/retrieve files
- **Data Management**: e.g. metadata catalogue, file transfer service

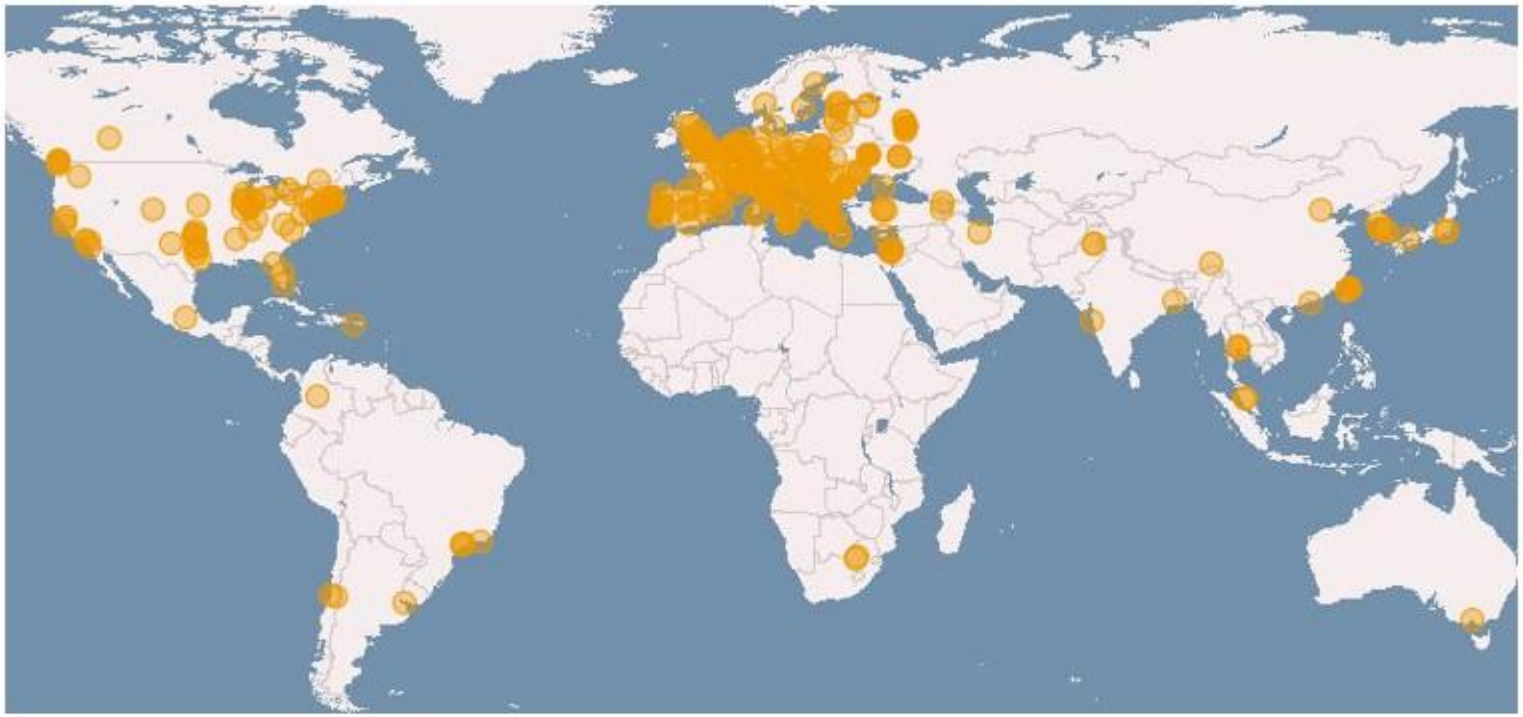
Based on **open standards** and **open source software**

Integrate heterogeneous infrastructure/technologies

Added Value

- **Uniform access** to distributed computing capabilities to run large-scale computational jobs processing big data and preventing single vendor lock-in
- Possibility to **federate your own resources**
- Facilitate **collaboration** across communities and borders by **sharing compute and data**

European Grid Infrastructure



Distributed and federated data and computing facilities
Grid and Cloud compute platform
340 data centres in 34 National Grid Initiatives/EIROs
435,000 logical CPU cores

10 years of support to science
> 200 research projects
190 PB disk, 180 PB tape
1.6 M job/day
> 99.6% reliability

Slides courtesy of EGI

Federated Cloud

Infrastructure to deploy on-demand IT services for managing and processing your research data?

European federation of publicly-funded community clouds
- **Cloud Compute**: deploy/manage virtual machines (VM)
- **Cloud Storage**: store/access/retrieve digital objects incl. metadata
- **Cloud Marketplace**: store/retrieve public & private VM image lists
- **Image Distribution**: CSPs integration for automated local updates

Based on **open standards** and **open source software**

Integrate heterogeneous cloud technologies
- OpenStack, OpenNebula, CloudStack,…

Integrate with commercial providers
- Within EGI or through the Helix Nebula Marketplace

Added Value

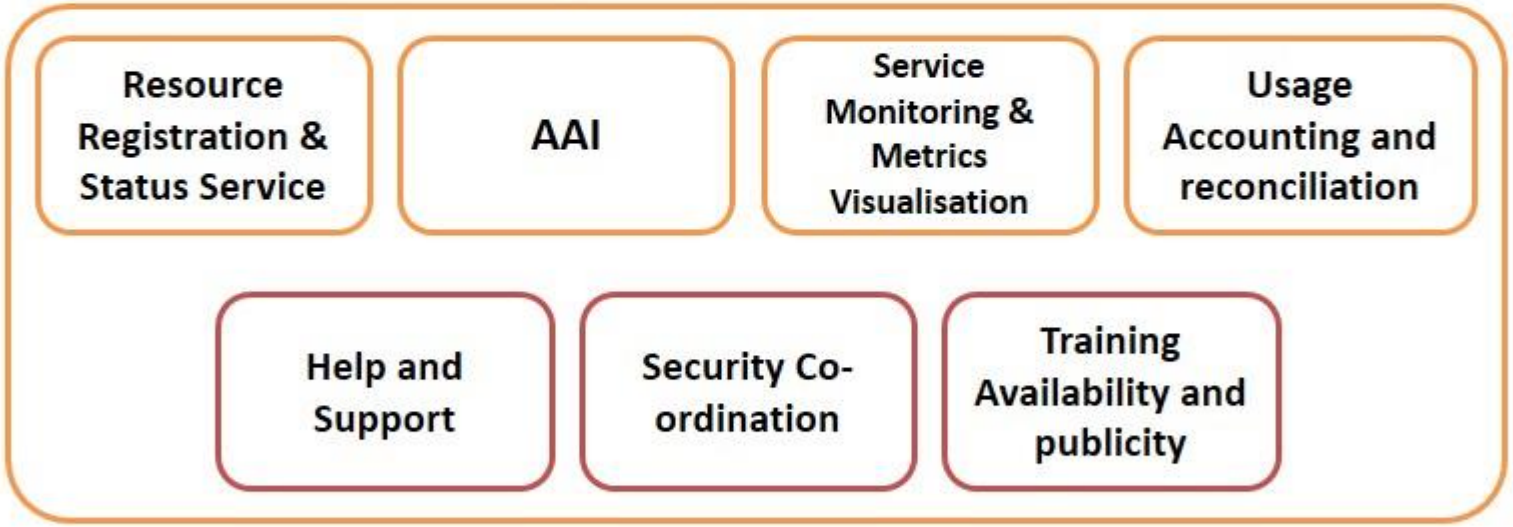
- Uniform access, **no lock-in** and **on-demand scale out** capabilities
- Easy deployment of **own/customised services**
- Possibility to **federate** existing institutional clouds
- **Efficiency** by **co-locating big data + cloud computing**

Launch into production: May 2014

Slides courtesy of EGI

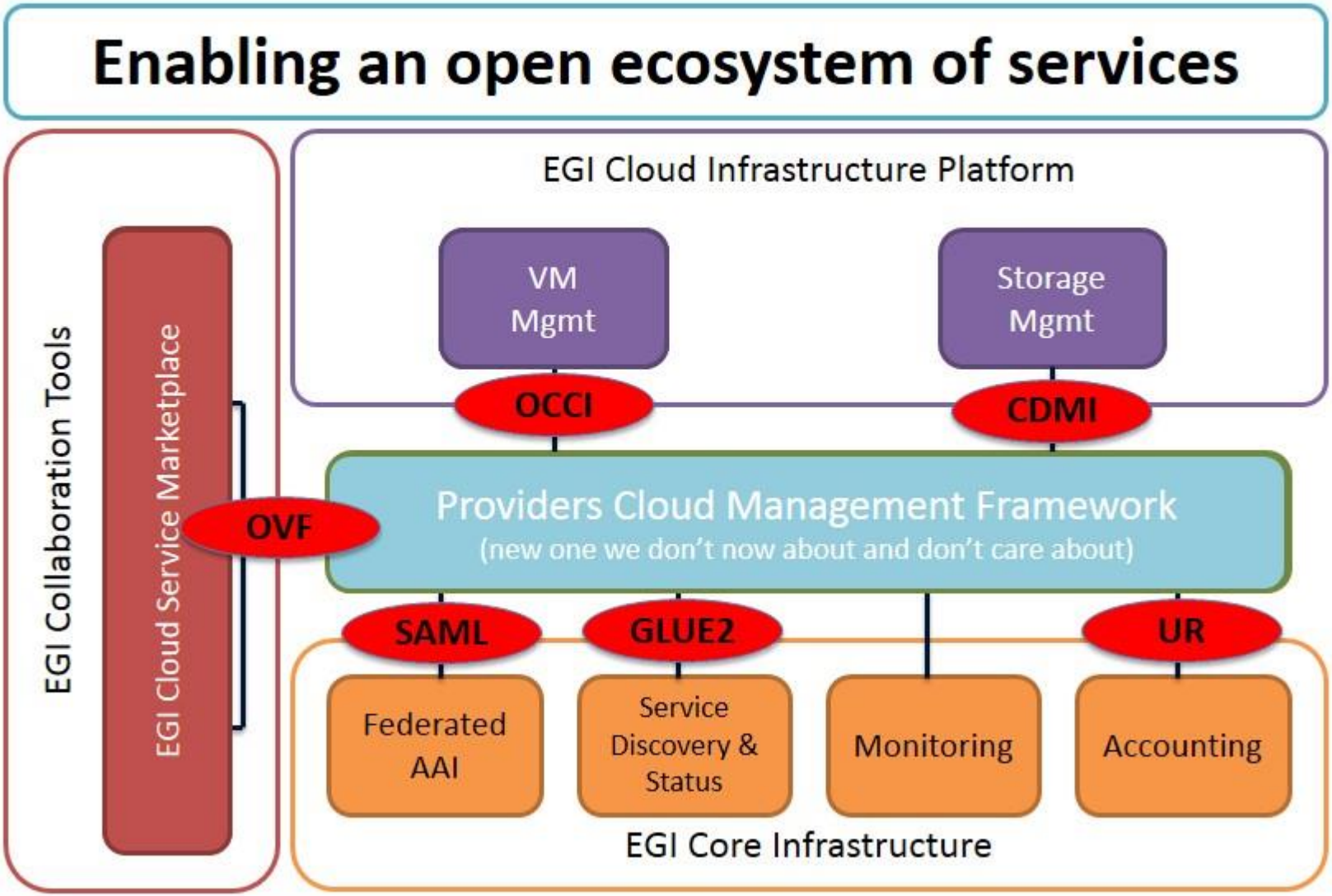
EGI Core Platform

Services that federate and integrate any user facing functional service deployed in **production**



For e-Infrastructures & Research Infrastructures

EGI's Cloud Infrastructure



Slides courtesy of EGI

EGI's Cloud Vision

To support the digital European Research Area through a pan-European research infrastructure based on an open federation of reliable services that provide uniform access to computing and data resources provided by the public and private sector.

EGI federated Cloud capability vision

10M cores Cloud compute

1 EB Cloud storage

- Paving the way for a federated cloud in Europe
  - Full production, May 2014 – **5,000** cores, **225 TB** storage
  - End of year 2014 (planned) – **18,000** cores, **6000 TB** storage
    (3.6x) **(26x)**

EGI Services for Federated Operations

- Activities and tools for the operations of distributed services
  - Central operations tools (message brokers, operations dashboards, VO management, service and security monitoring, service registry)
  - Federated accounting (distributed repositories and portal)
  - Technical support and incident management
  - Security operations coordination, policy development, software vulnerability
  - Software distribution, verification, validation

EGI Long-term vision for European RIs and ERA

- One European High Throughput Computing (HTC) and Cloud infrastructure
  - Technical integration
    - Europe – e.g. EUDAT, PRACE
    - World-wide (liaison) – e.g. OSDC, XSEDE, OSG, SAGrid, PIRE
  - Complemented with commercial (Cloud) Service Providers
- Distributed network of Competence Centres
  - Discipline / domain oriented
    - E.g. structural biology, Astronomy, Archeology
  - Cross-cutting competence centres
    - E.g. security, Cloud Compouting, parallel computing, **Big Data**

SDI and Cloud Computing

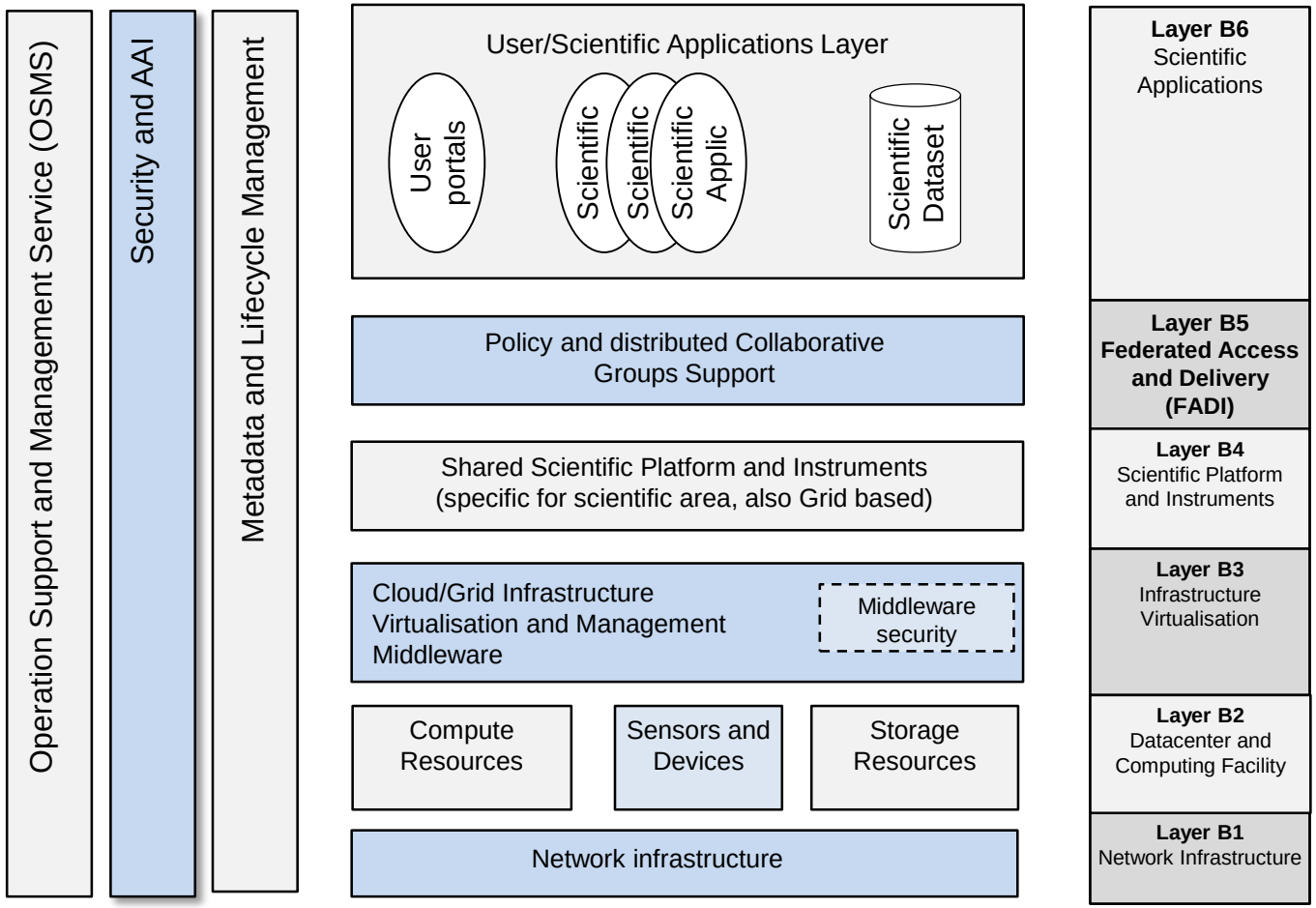
General requirements to SDI for emerging Big Data Science

- Support for *long running experiments and large data volumes* generated at high speed
- *Multi-tier inter-linked data distribution and replication*
- *On-demand infrastructure provisioning* to support data sets and scientific workflows, mobility of data-centric scientific applications
- Support of *virtual scientists communities*, addressing dynamic user groups creation and management, federated identity management
- Support for the *whole data lifecycle* including metadata and data source linkage
- *Trusted environment* for data storage and processing
  - Research need to trust SDI to put all their data on it
- Support for data integrity, confidentiality, accountability
- *Policy binding to data* to protect privacy, confidentiality and IPR

Defining Architecture framework for SDI and FADI

- Scientific Data Lifecycle Management (SDLM) model
- e-SDI multi-layer architecture model
- Capabilities, Roles, Actors
  - RORA (Resource-Ownership-Role-Actor) model defines relationship between resources, owners, managers, users
  - Initially defined for telecom domain
  - Potentially new actor in SDI – Subject of data (e.g. patient, or scientific object/paper)
- Security and Federated Access Control and Delivery Infrastructure (FADI)
  - Authentication, Authorisation, Accounting
    - Federated Access Control and Identity Management
  - Extended to support data access control and operations on data
  - Trust management infrastructure

SDI Architecture Model and Federated Infrastructure components



Technologies and solutions

Scientific specialist applications
Library resources

FADI: Optical Network Infrastructure
Federated Identity Management: eduGAIN, REFEDS, VOMS, InCommon, **OCX**

PRACE/DEISA

Grid/Cloud, **OCX**

Clouds

Autobahn, **eduroam**

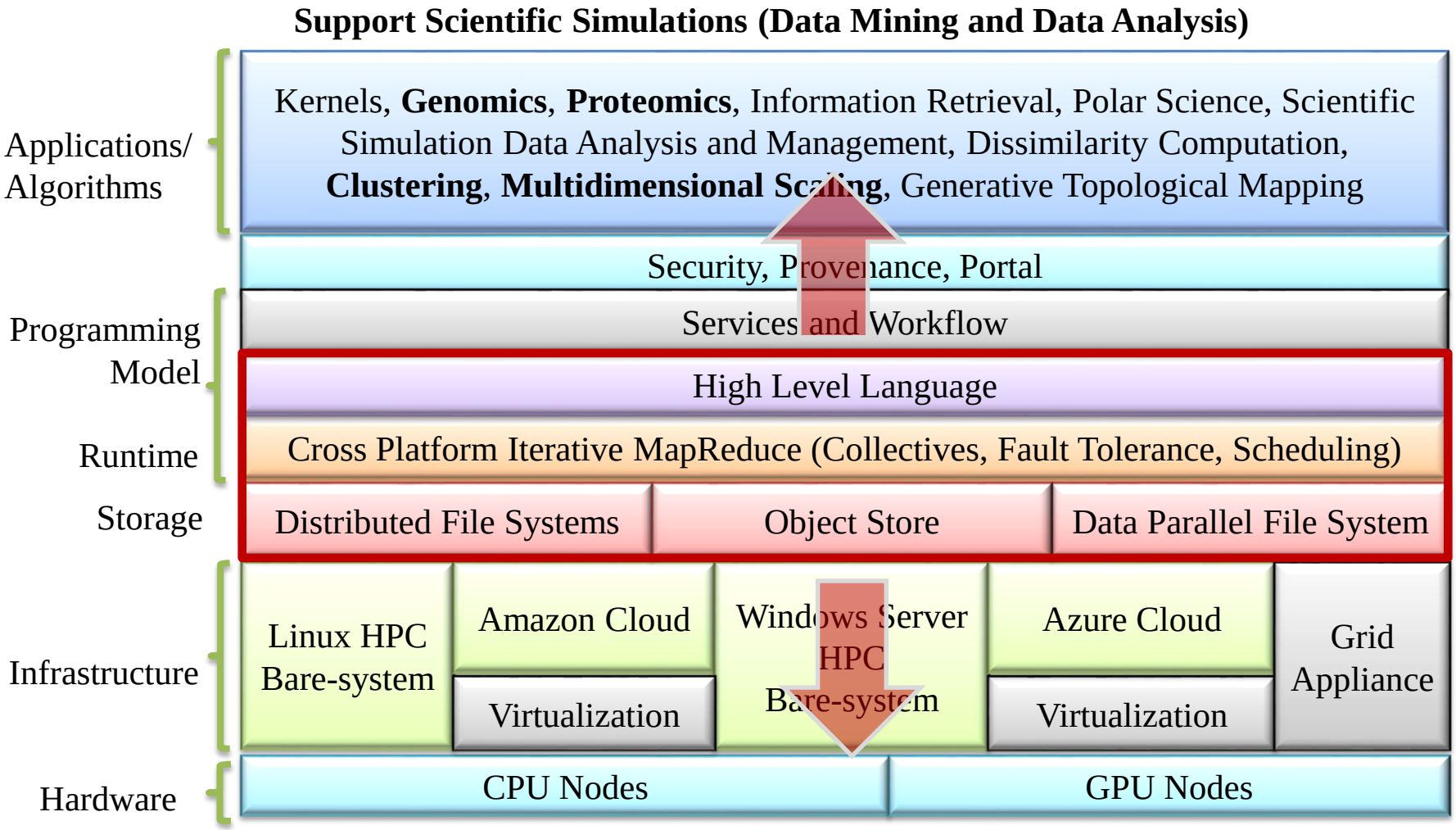
SDI Architecture Layers

- **Layer D1**: Network infrastructure layer represented by the general purpose Internet infrastructure and dedicated network infrastructure
- **Layer D2**: Datacenters and computing resources/facilities, including sensor network
- **Layer D3**: Infrastructure virtualisation layer that is represented by the Cloud/Grid infrastructure services and middleware supporting specialised scientific platforms deployment and operation
- **Layer D4**: (Shared) Scientific platforms and instruments specific for different research areas
- **Layer D5**: Federated Access and Delivery Infrastructure: Federation infrastructure components, including policy and collaborative user groups support functionality
- **Layer D6**: Scientific applications and user portals/clients

SDI move to Clouds

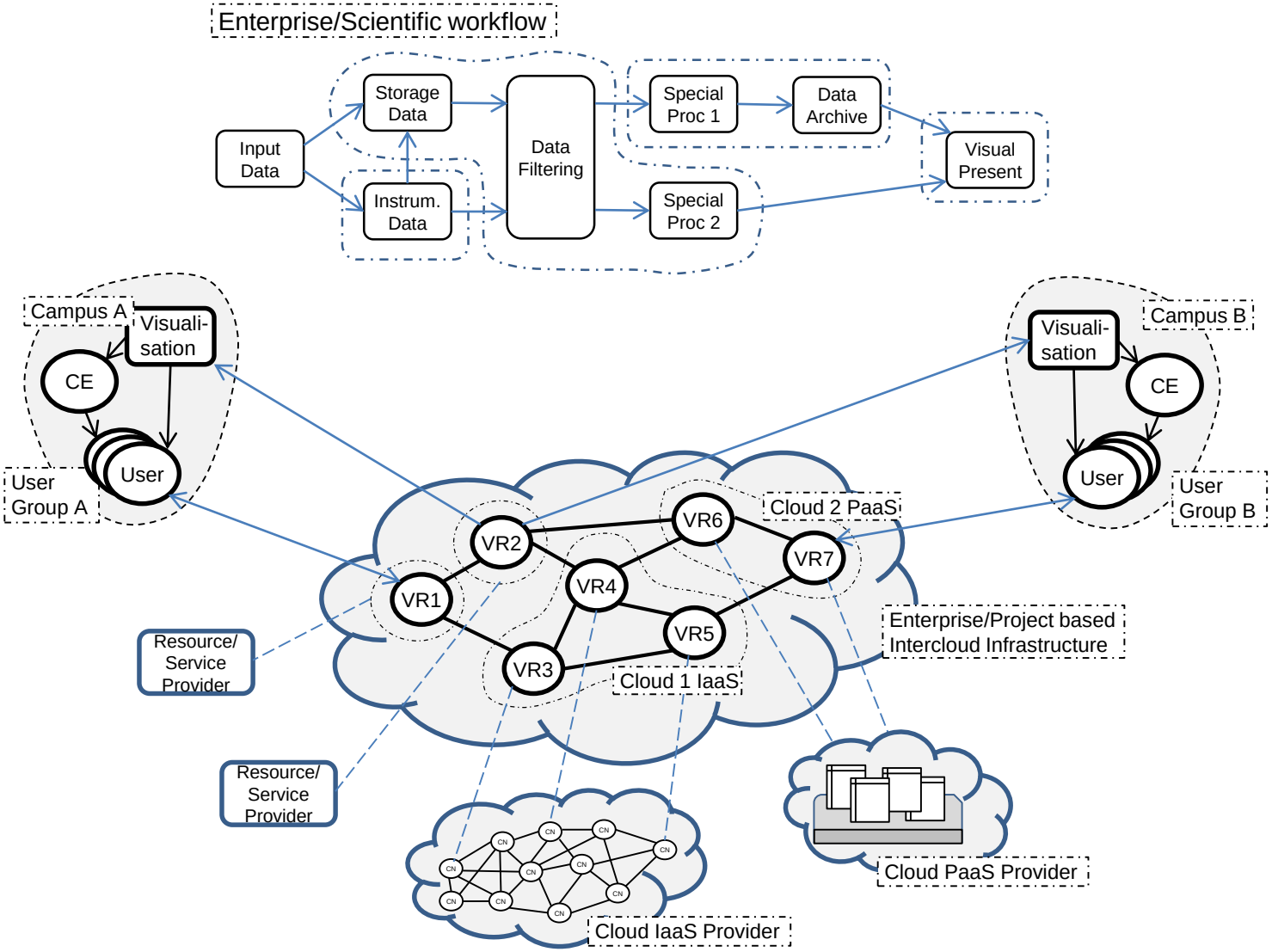
- Cloud technologies allow for infrastructure virtualisation and its profiling for specific data structures or to support specific scientific workflows
  - Clouds provide just right technology for infrastructure virtualisation to support data sets
  - *Complex distributed data require infrastructure*
    - *Demand for inter-cloud infrastructure*
- Cloud can provide infrastructure on-demand to support project related scientific workflows
  - Similar to Grid but with benefits of the full infrastructure provisioning on-demand
- Software Defined Infrastructure Services
  - As wider than currently emerging SDN (Software Defined Networks)
- Distributed Hadoop clusters for HPC and MPP

Data Analysis Architecture [ref]

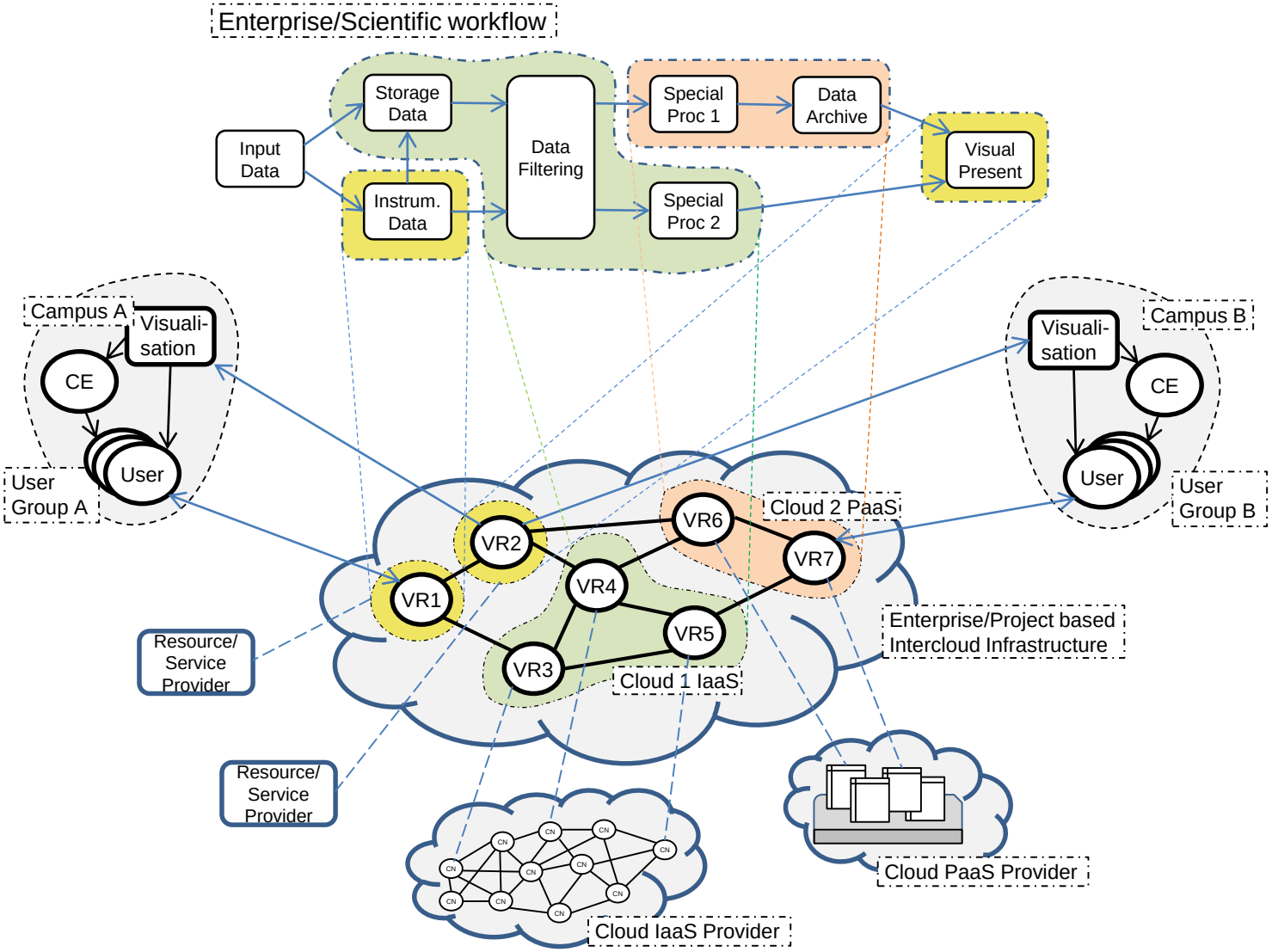


[ref] Source: presentation by Judy Qiu "Analysis Tools for Data Enabled Science" at the Big Data Analytics Workshop (BDAW2013)

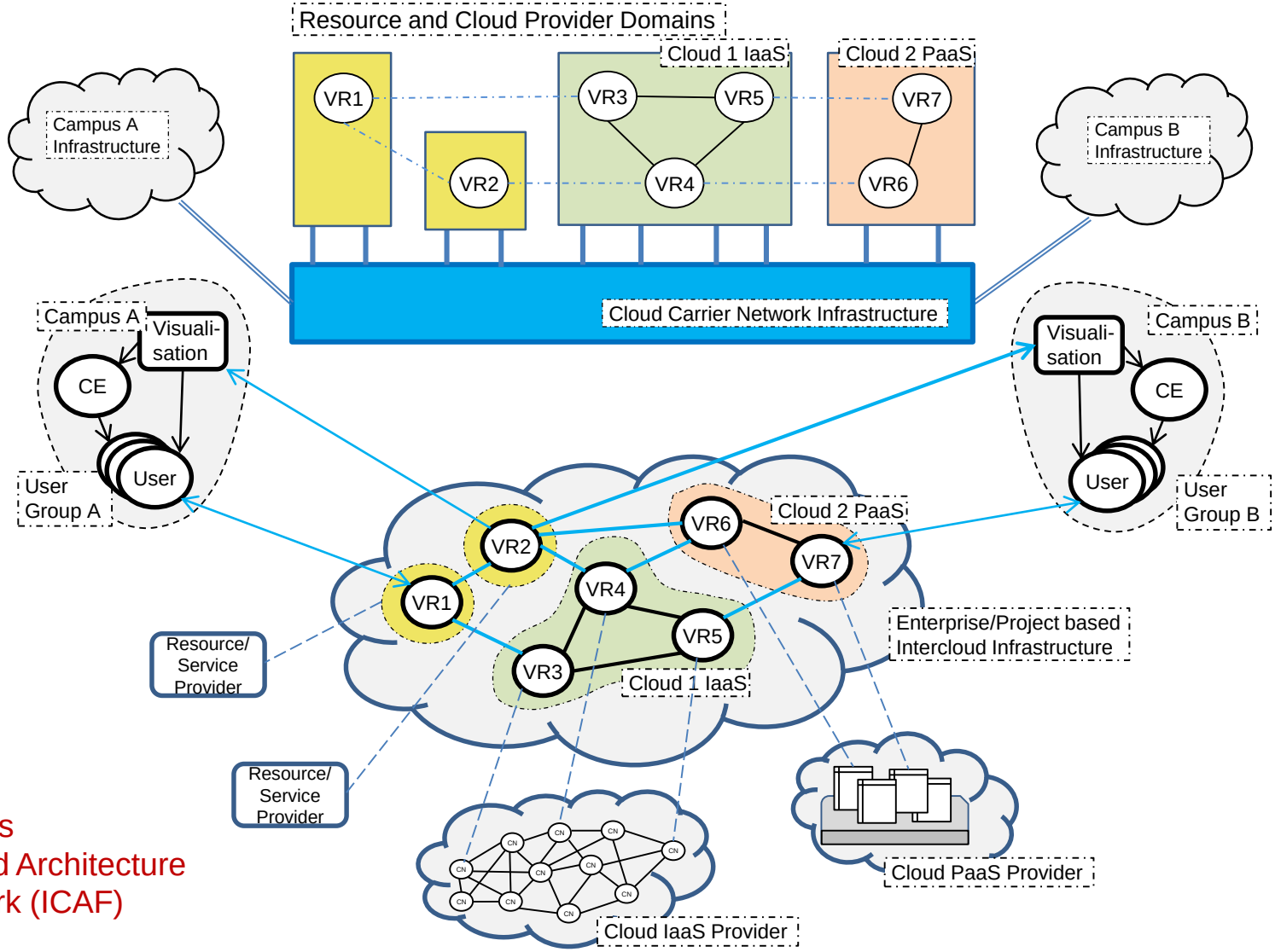
General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure



General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure



General use case for infrastructure provisioning: Logical Infrastructure => Network Infrastructure (1)



Defined as InterCloud Architecture Framework (ICAF)

InterCloud Architecture Framework (ICAF) Components (proposed by UvA, submitted to IETF)

- Multi-layer Cloud Services Model (CSM)
  - Combines IaaS, PaaS, SaaS into multi-layer model with inter-layer interfaces
  - Including interfaces between cloud service layers and virtualisation platform
- InterCloud Control and Management Plane (ICCMP)
  - Allows signaling, monitoring, dynamic configuration and synchronisation of the distributed heterogeneous clouds
  - Including management interface from applications to network infrastructure and virtualisation platform
- InterCloud Federation Framework (ICFF)
  - Defines set of protocols and mechanisms to ensure heterogeneous clouds integration at service and business level
  - Addresses Identity Federation, federated network access, etc.
- InterCloud Operations Framework (ICOF)
  - RORA model: Resource, Ownership, Role, Action
  - Business processes support, cloud broker and federation operation

Intercloud Architecture for Interoperability and Integration, Release 1, Draft Version 0.5. SNE Technical Report 2012-03-02, 6 September 2012
http://staff.science.uva.nl/~demch/worksinprogress/sne2012-techreport-12-05-intercloud-architecture-draft05.pdf

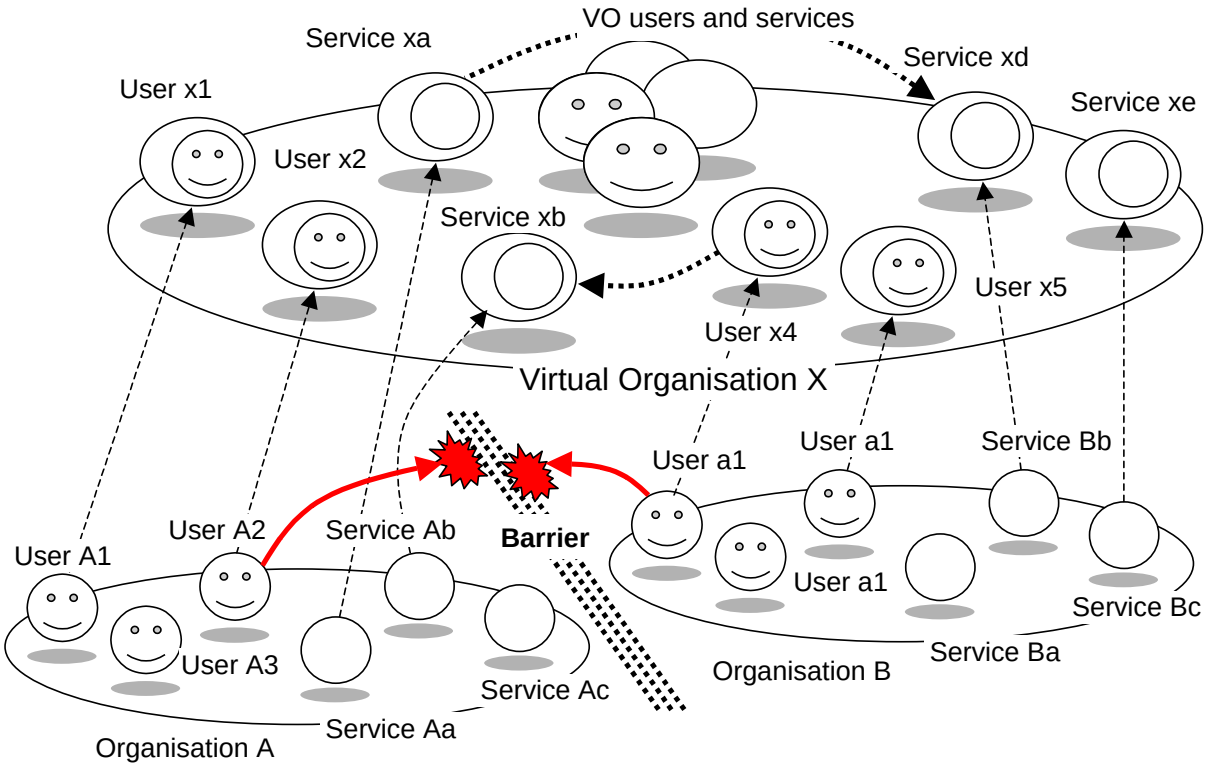
Cloud Federation and Federated AAI

- Virtual Organisations legacy Federation model
- Users and resources federation in clouds
  - Federation models
- Federated Access Control in clouds

Cloud Federation and VO based Federated Grid Infrastructure

- Grid federates resources and users by creating Virtual Organisations (VO)
  - VO membership is maintained by assigning VO membership attributes to VO resources and members
    - VO Membership Service (VOMS)
  - Users remain members of their Home Organisations (HO)
    - AuthN takes place at HO or Grid portal
    - To access VO resources, VO members need to obtain VOMS certificate or VOMS credentials
  - Resources remain under control of the resource owner organisation Grid Centers
  - Scalability and on-demand provisioning issues
- In clouds, both resources and user accounts are created/provisioned on-demand as virtualised components/entities
  - User accounts/identities can be provisioned together with access rights to virtual resources

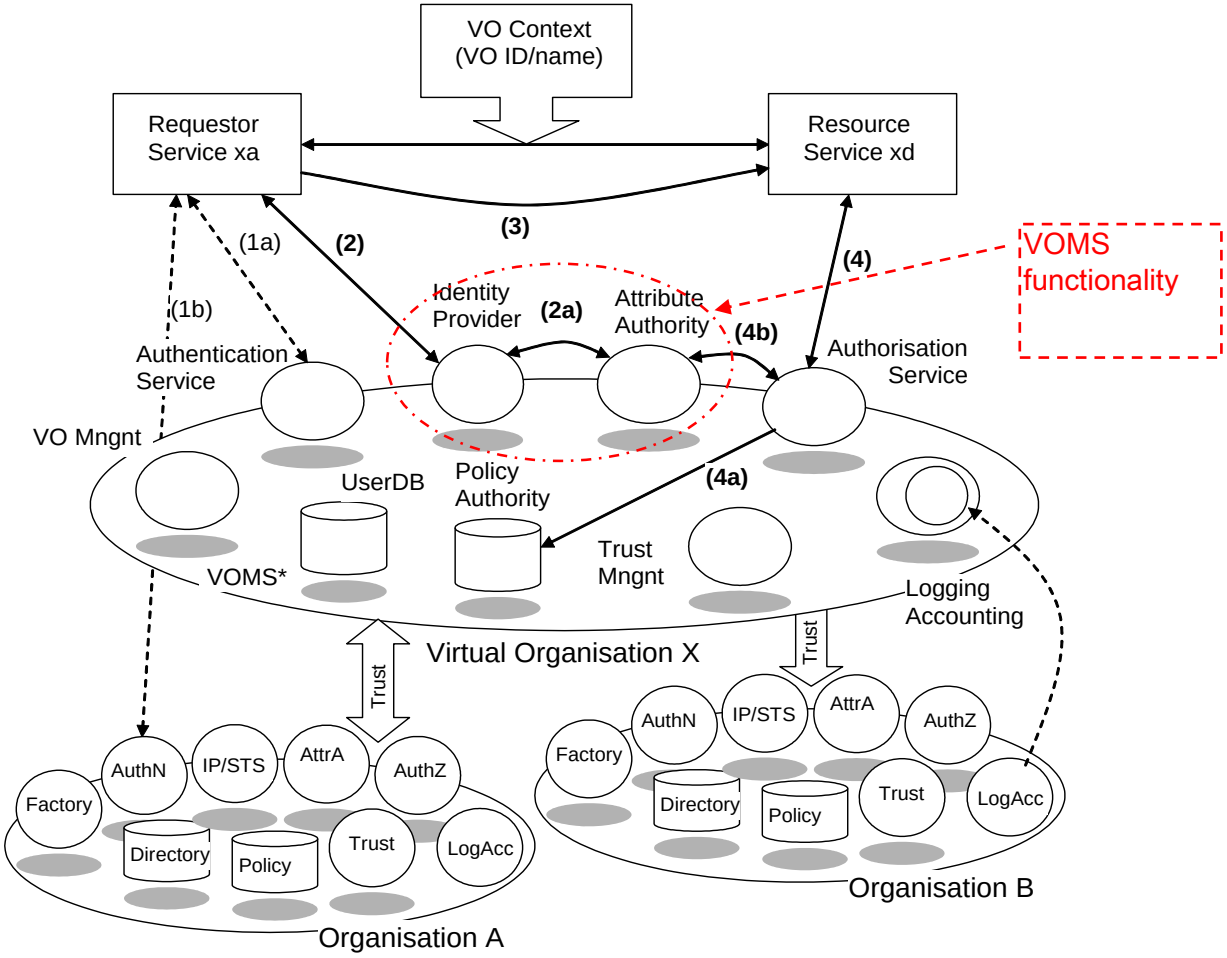
VO bridging inter-organisational barriers



- VO allows bridging inter-organisational barriers without changing local policies
  - Requires VO Agreement and VO Security policy
  - VO dynamics depends on implementation but all current implementations are rather static

VO-based Dynamic Security Associations in Collaborative Grid Environment, COLSEC'06 Workshop, 15 May 2006, Las Vegas

Example VO Security services operation



VO-based Dynamic Security Associations in Collaborative Grid Environment, COLSEC'06 Workshop, 15 May 2006, Las Vegas

Cloud Federation: (new) Actors and Roles

- Cloud Service Provider (CSP)
- Cloud Customer (organisational)
  - Multi-tenancy is provided by virtualisation of cloud resources provided to all/multiple customers
  - Cloud tenant is associated with the customer organisation
- Cloud User (end user)
  - Cloud User can be a user/role for different tenants/services
- Cloud (Service) Broker
- Identity Provider (IDP)
- Cloud Carrier
- Cloud Service Operator
- Cloud Auditor

Cloud Federation – Scaling up and down

- Scalability is one of the main cloud feature
  - To be considered in the context of hybrid cloud service model
    - Cloud burst and outsourcing enterprise services to cloud
    - Cloud services migration and replication between CSP
- Scaling up
  - Identities provisioning
  - Populating sessions context
- Scaling down
  - Identity deprovisioning: Credentials revocation?
  - Sessions invalidation vs restarting
- Initiated by provider and by user/customer

Cloud Federation Models – Identified models

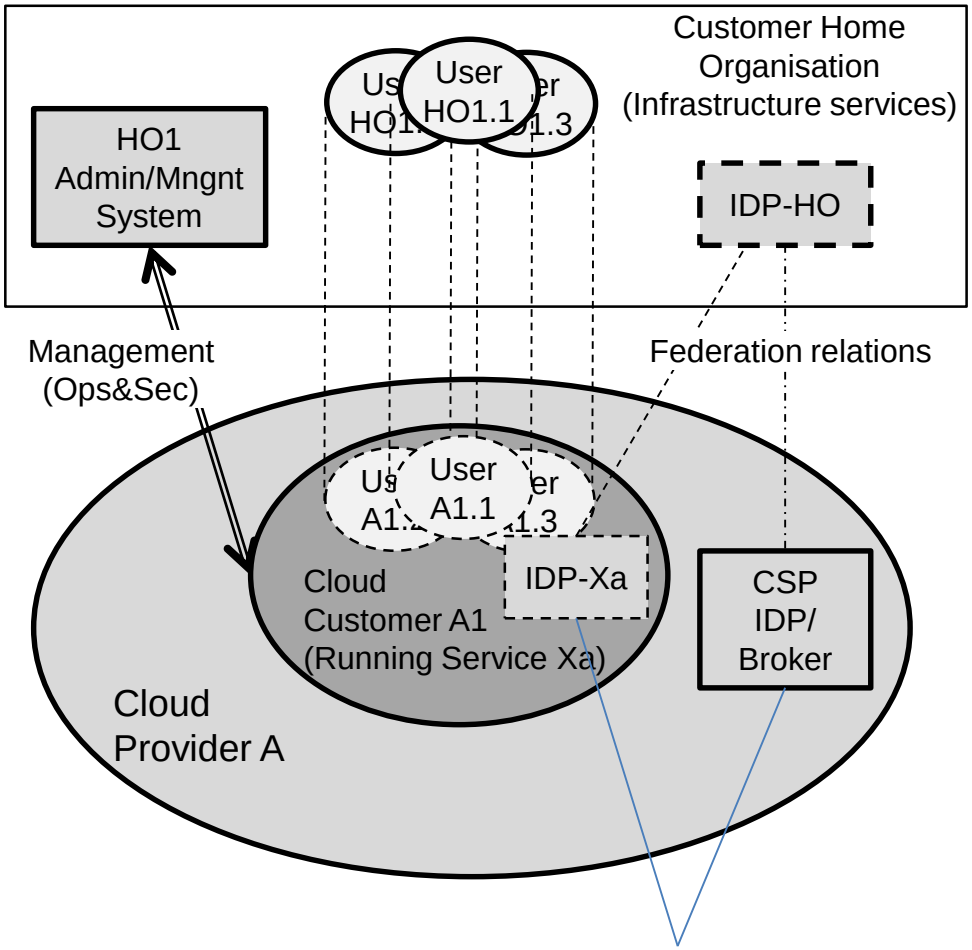
User/customer side federation

- (1.1) Federating users/HO and CSP/cloud domains
  - Customer doesn't have own IDP (IDP-HO)
  - Cloud Provider's IDP is used (IDP-CSP)
- (1.2) Federating HO and CSP domains
  - Customer has own IDP-HO1
  - It needs to federate with IDP-CSP, i.e. have ability to use HO identities at CSP services
- (1.3) Using 3rd party IDP for external users
  - Example: Web server is run on cloud and external user are registered for services

Provider (resources) side federation

- (2.1) Federating CSP's/multi-provider cloud resources
  - Used to outsource and share resources between CSP
  - Typical for community clouds

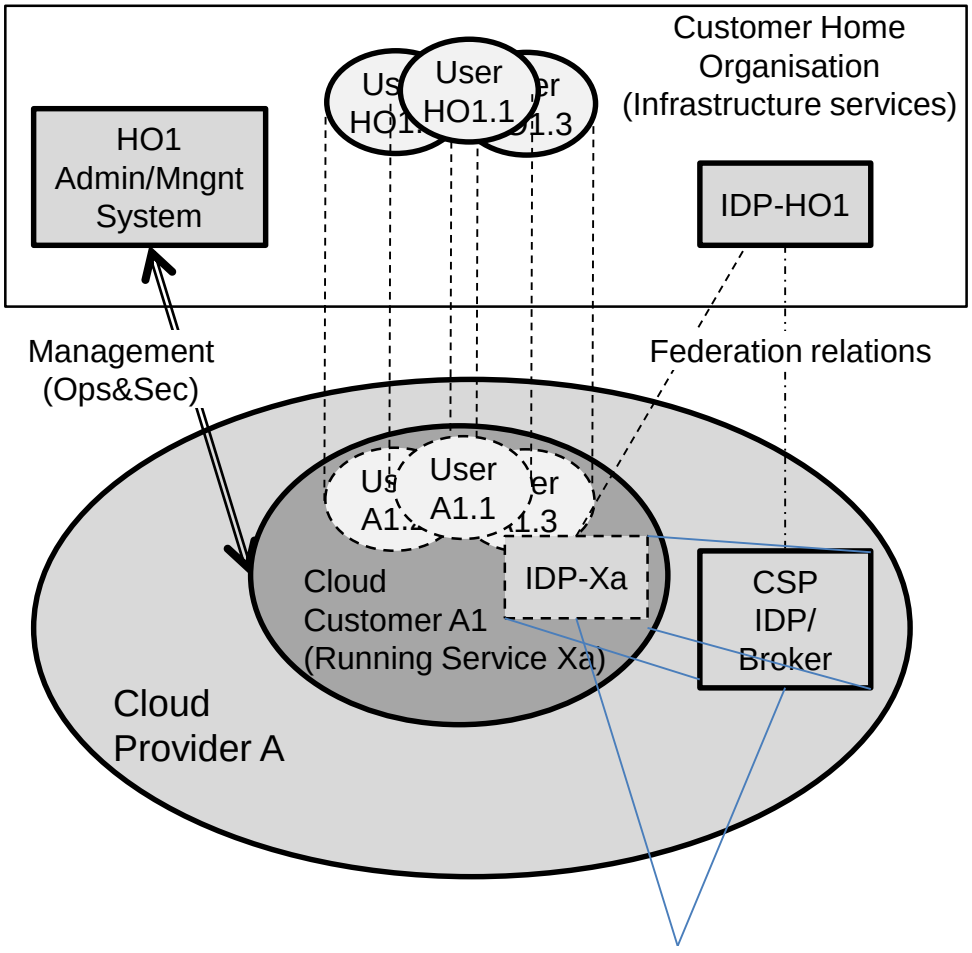
Basic Cloud Federation model (1.1) – Federating users/HO and CSP/cloud domains (no IDP-HO)



User side Federation

- Simple/basic scenario 1: Federating Home Organisation (HO) and Cloud Service Provider (CSP) domains
- Cloud based services created for users from HO1 and managed by HO1 Admin/Management system
- Involved major actors and roles
  - CSP – Customer – User
  - IDP/Broker
- Cloud accounts A1.1-3 are provisioned for each user 1-3 from HO with 2 options
  - Individual accounts with new ID::pswd
  - Mapped/federated accounts that allows SSO/login with user HO ID::pswd
- Federated accounts may use Cloud IDP/Broker (e.g. KeyStone) or those created for Service Xa

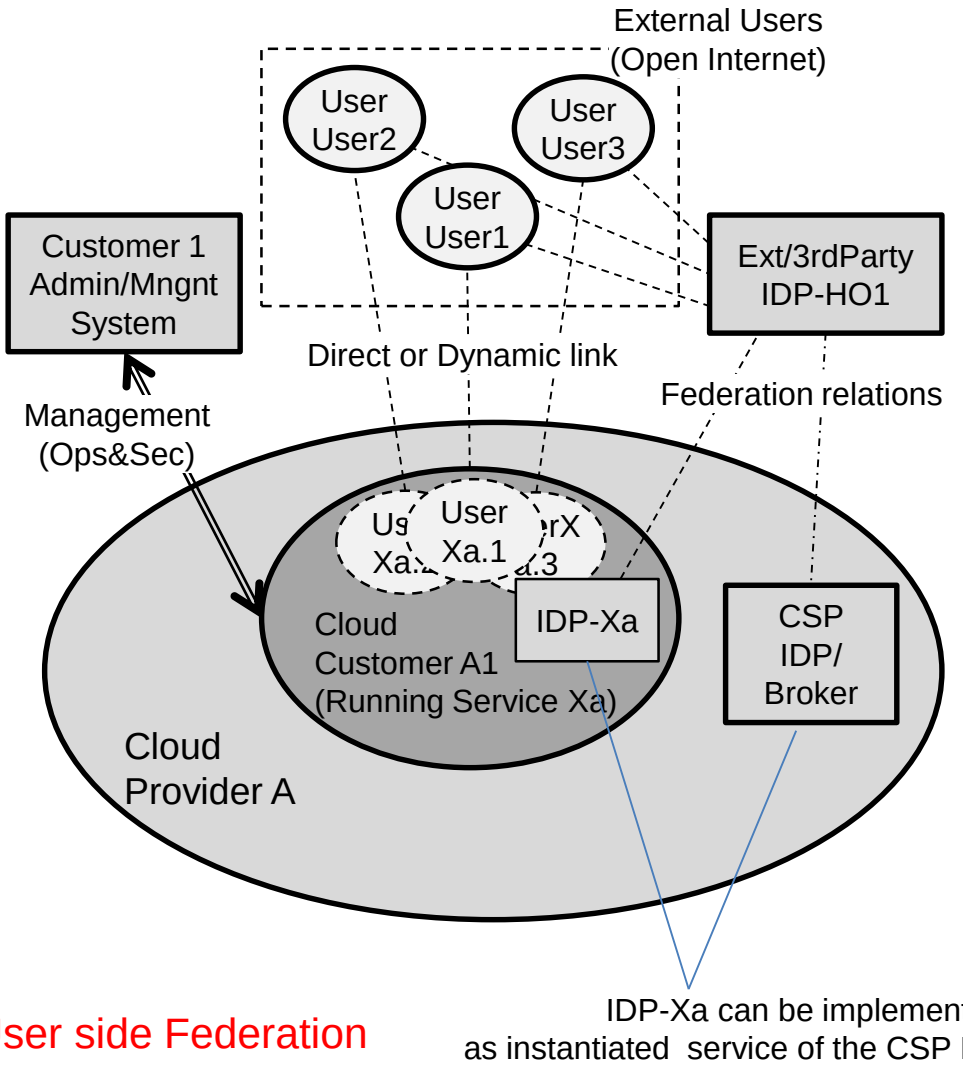
Basic Cloud Federation model (1.2) – Federating HO and CSP domains (IDP-HO1 and IDP-CSP)



User side Federation

- Simple/basic scenario 1: Federating Home Organisation (HO) and Cloud Service Provider (CSP) domains
- Cloud based services created for users from HO1 and managed by HO1 Admin/Management system
- Involved major actors and roles
  - CSP – Customer – User
  - IDP/Broker
- Cloud accounts A1.1-3 are provisioned for each user 1-3 from HO with 2 options
  - Individual accounts with new ID::pswd
  - Mapped/federated accounts that allows SSO/login with user HO ID::pswd
- Federated accounts may use Cloud IDP/Broker (e.g. KeyStone) or those created for Service Xa

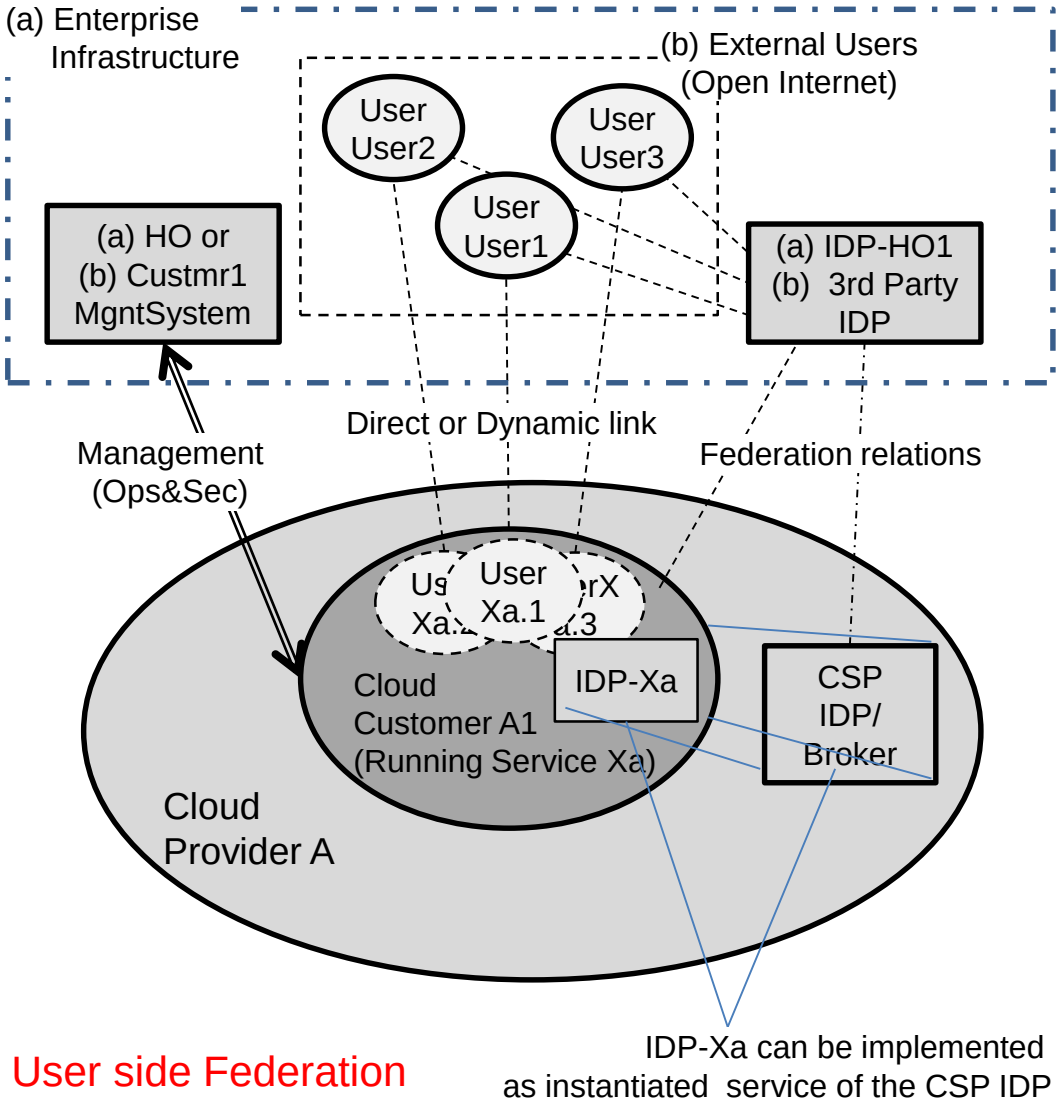
Basic Cloud Federation model (1.3) – Using 3rd party IDP for external users



User side Federation

- Simple/basic scenario 2: Federating Home Organisation (HO) and Cloud Service Provider (CSP) domains
- Cloud based services created for external users (e.g. website) and managed by Customer 1
- Involved major actors and roles
  - CSP – Customer – User
  - IDP/Broker
- Cloud accounts A1.1-3 are provisioned for each user 1-3 from HO with 2 options
  - Individual accounts with new ID::pswd
  - Mapped/federated accounts that allows SSO/login with user HO ID::pswd
- Federated accounts may use Cloud IDP/Broker (e.g. KeyStone) or those IDP-Xa created for Service Xa

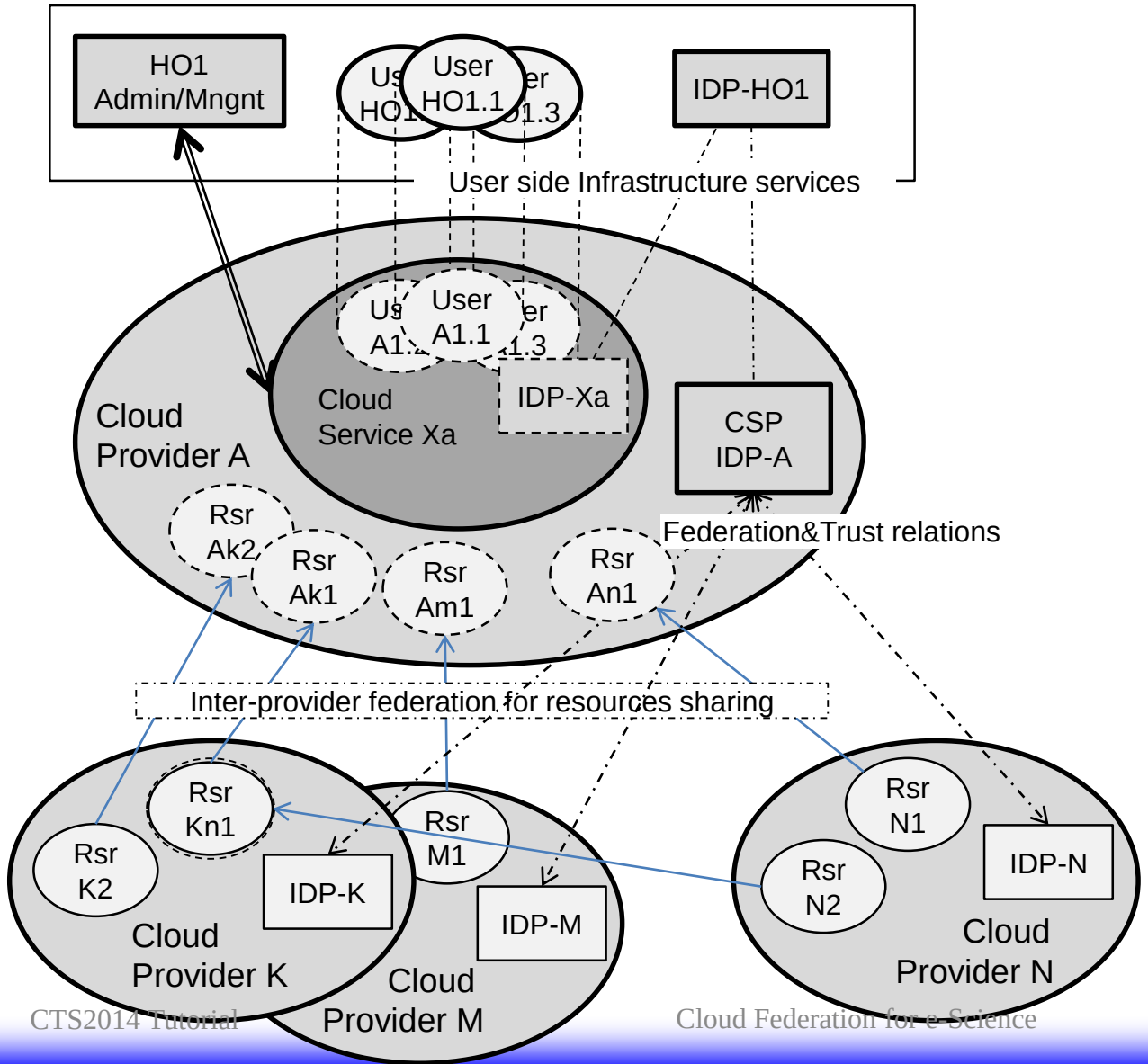
Basic Cloud Federation model – Combined User side federation



User side Federation

- Simple/basic scenario 2: Federating Home Organisation (HO) and Cloud Service Provider (CSP) domains
- Cloud based services created for external users (e.g. website) and managed by Customer 1
- Involved major actors and roles
  - CSP – Customer – User
  - IDP/Broker
- Cloud accounts A1.1-3 are provisioned for each user 1-3 from HO with 2 options
  - Individual accounts with new ID::pswd
  - Mapped/federated accounts that allows SSO/login with user HO ID::pswd
- Federated accounts may use Cloud IDP/Broker (e.g. KeyStone) or those IDP-Xa created for Service Xa

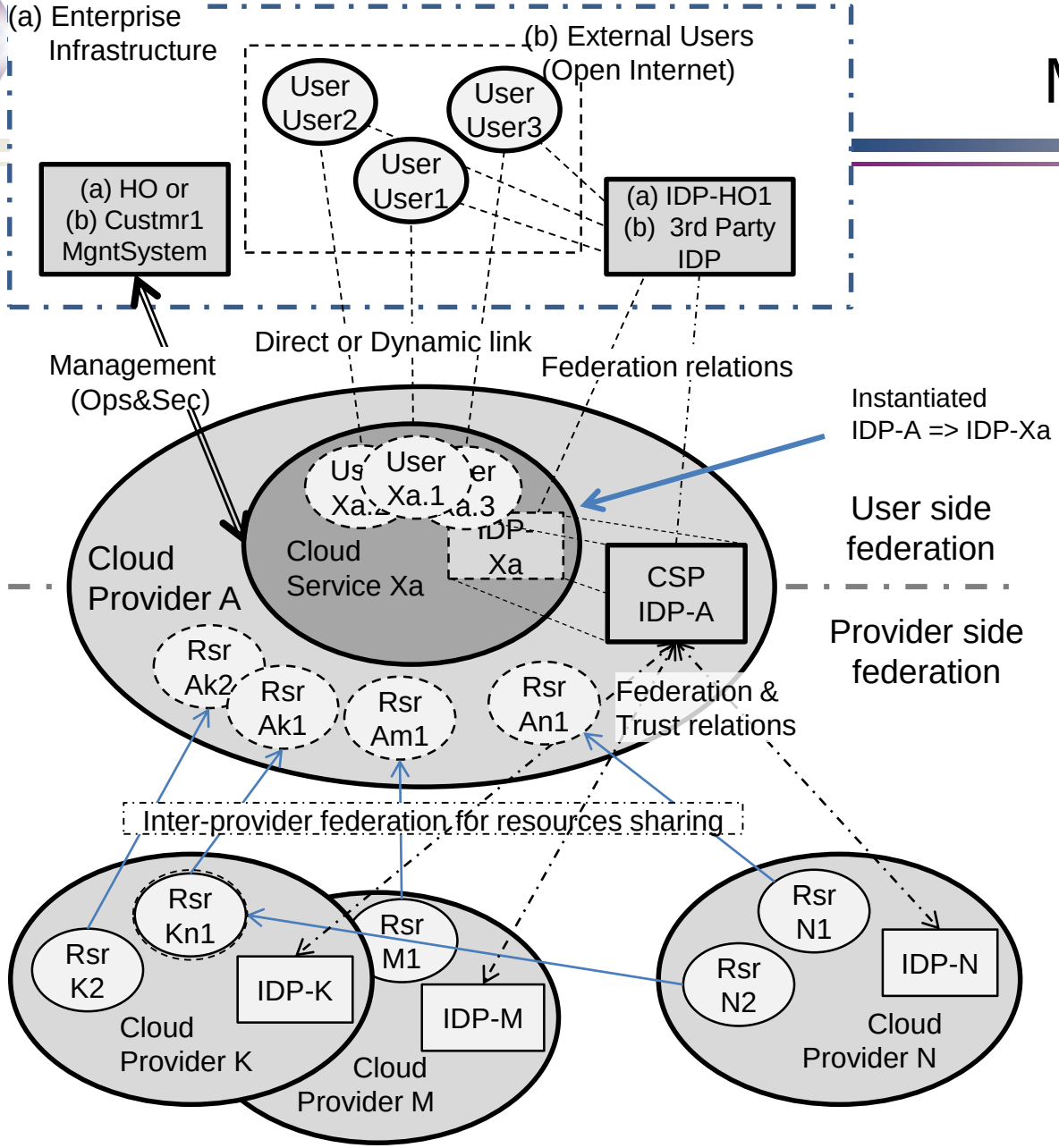
Basic Cloud Federation model (2.1) – Federating CSP's/multi-provider cloud resources



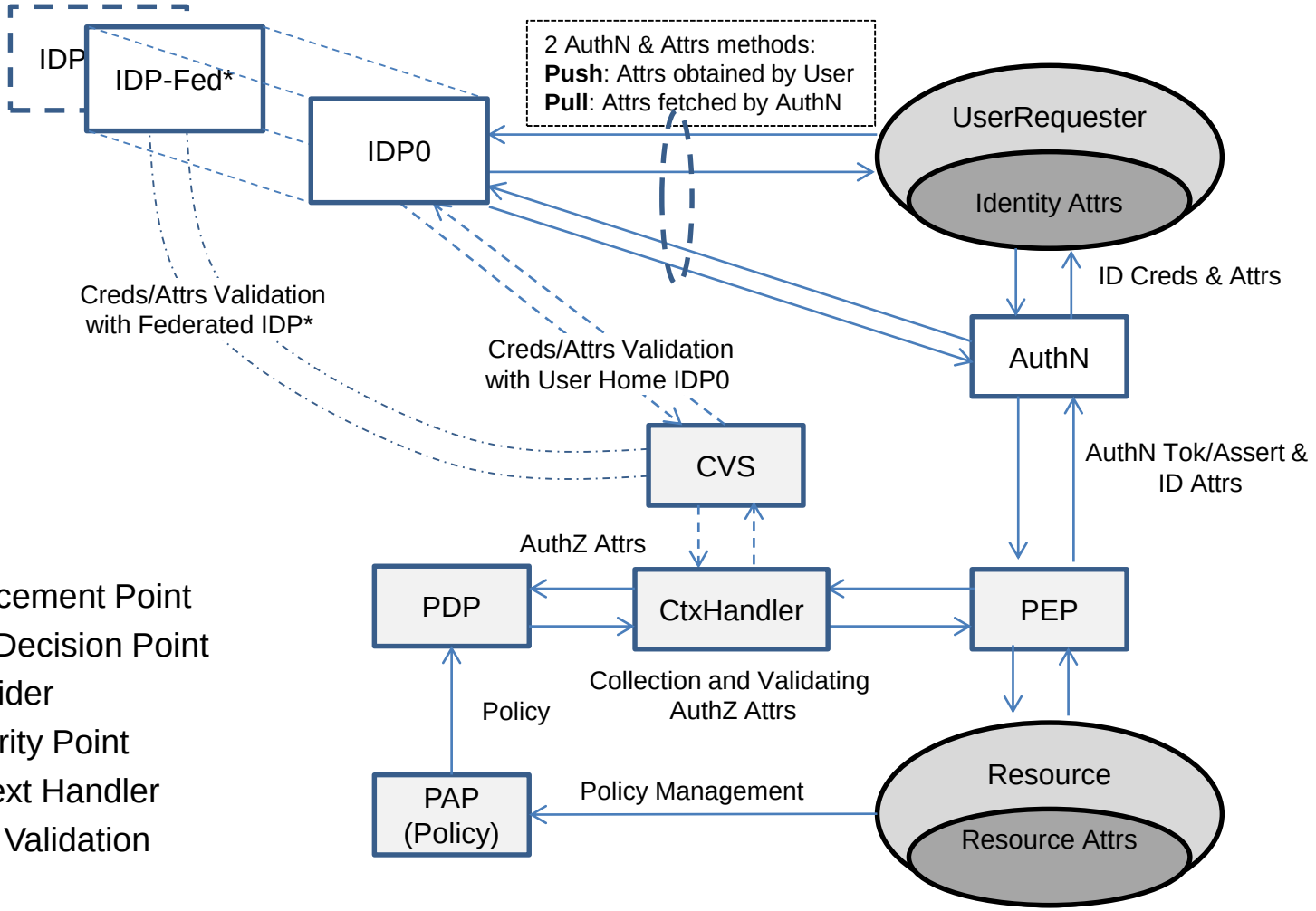
- Cloud provider side federation for resources sharing
- Federation and Trust relations are established between CSP's via Identity management services, e.g. Identity Providers (IDP)
  - May be bilateral or via 3rd party/broker service
- Includes translation or brokering
  - Trust relations
  - Namespaces
  - Attributes semantics
  - Policies
- Inter-provider federation is transparent to customers/users

Provider side Federation

Cloud Federation Model - Combined

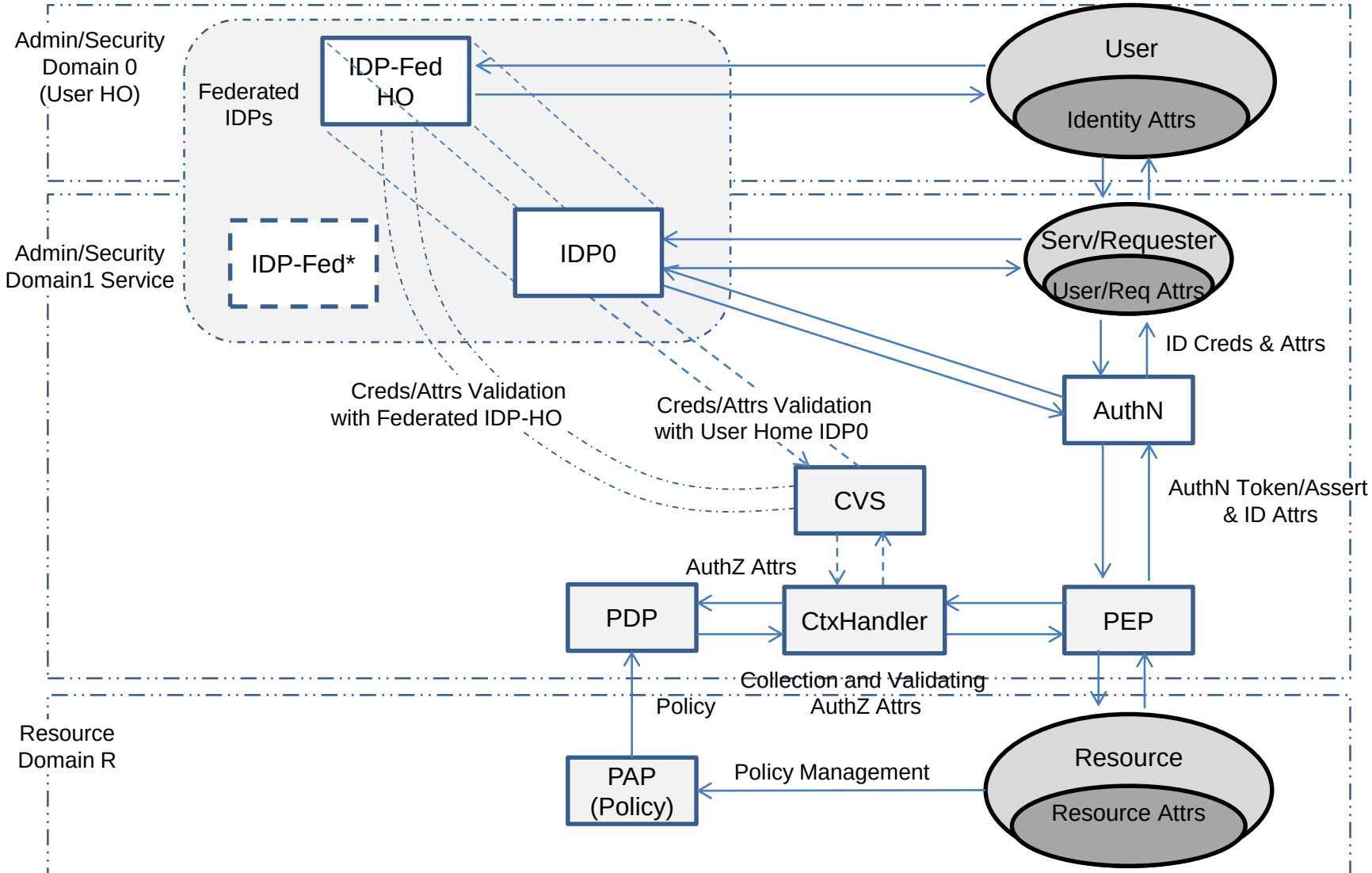


Basic AuthN and AuthZ services using Federated IDPs – For additional Credentials validation

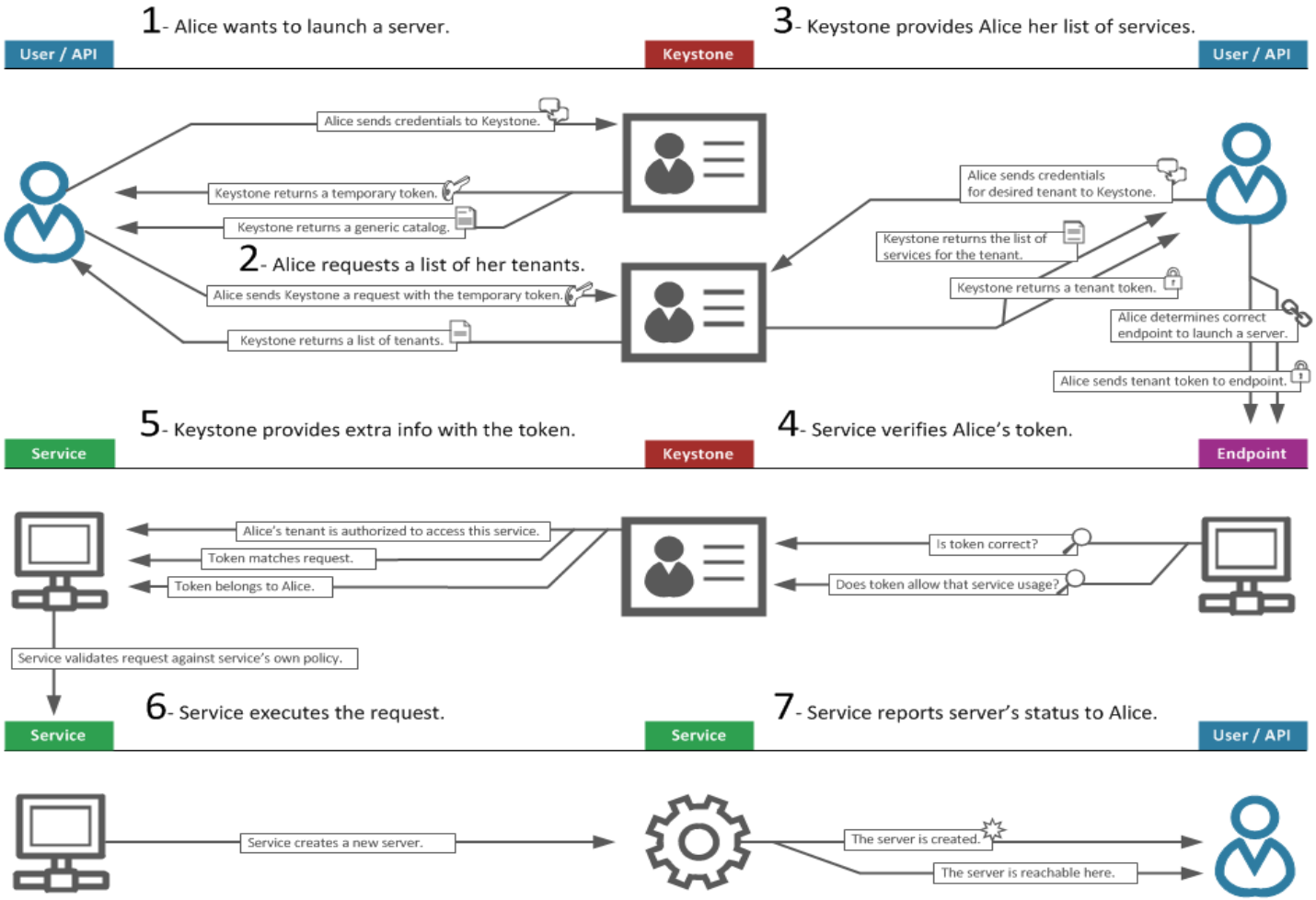


PEP - Policy Enforcement Point
PDP/ADF - Policy Decision Point
IDP – Identity Provider
PAP - Policy Authority Point
CtxHandler - Context Handler
CVS – Credentials Validation Service

Basic AuthN and AuthZ services using Federated IDPs – Federation/Trust domains



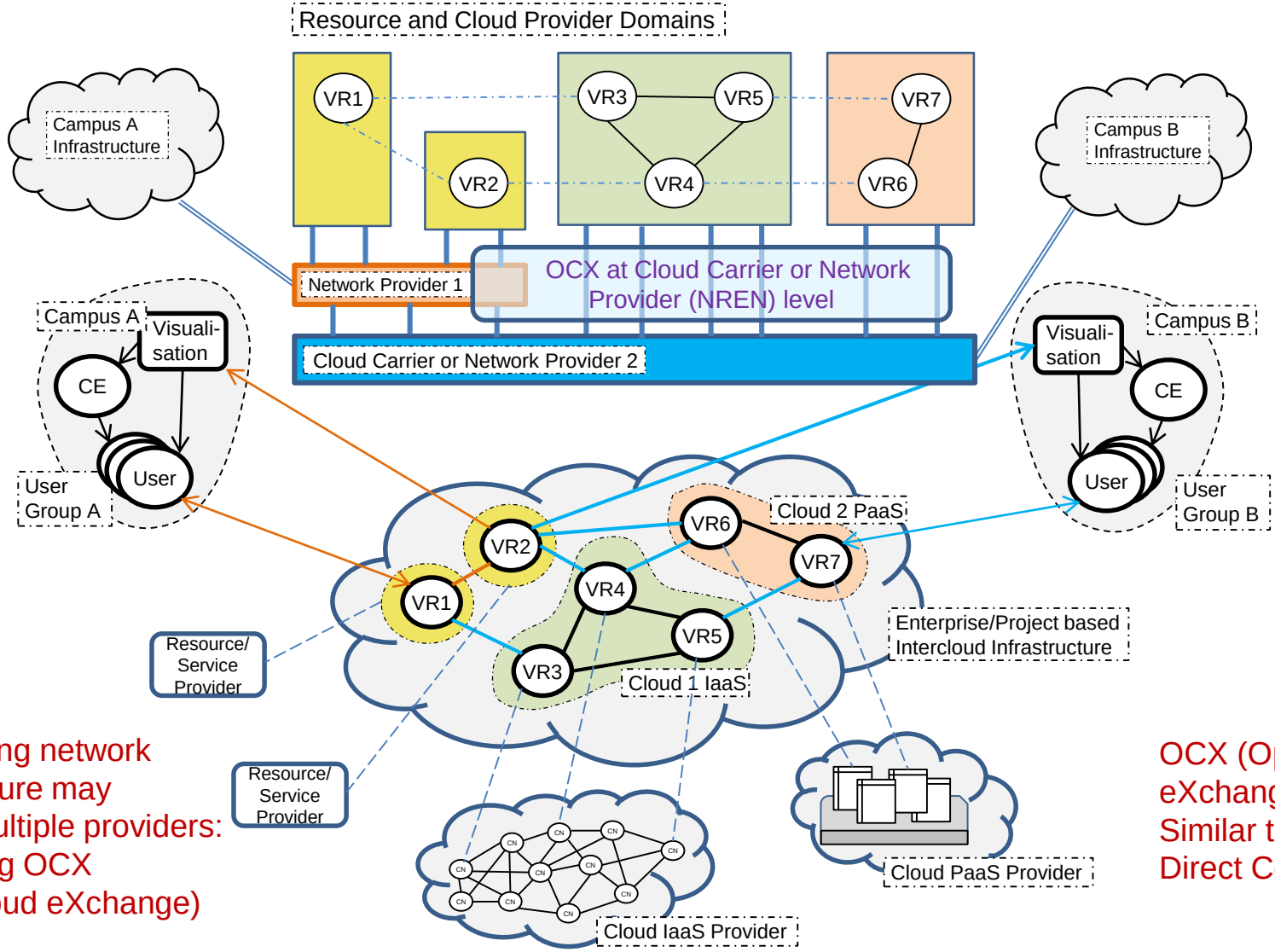
Implementation: Keystone Identity Server - Sequences



Implementation: Intercloud Federation Infrastructure and Open Cloud eXchange (OCX) in GEANT Infrastructure

- Open Cloud eXchange (OCX) initiative by GN3plus JRA1: Network Architectures for Horizon 2020
  - GEANT Network to support 2Tbps capacity backbone
  - SURFnet – PSNC 100 Gbps remote robotics demo at TNC2013
- From Software Defined Network (SDN) to Software Defined Infrastructure (SDI)
  - A new thinking beyond current challenges
- Federated Identity Management and Federated Access and Delivery Infrastructure (FADI)

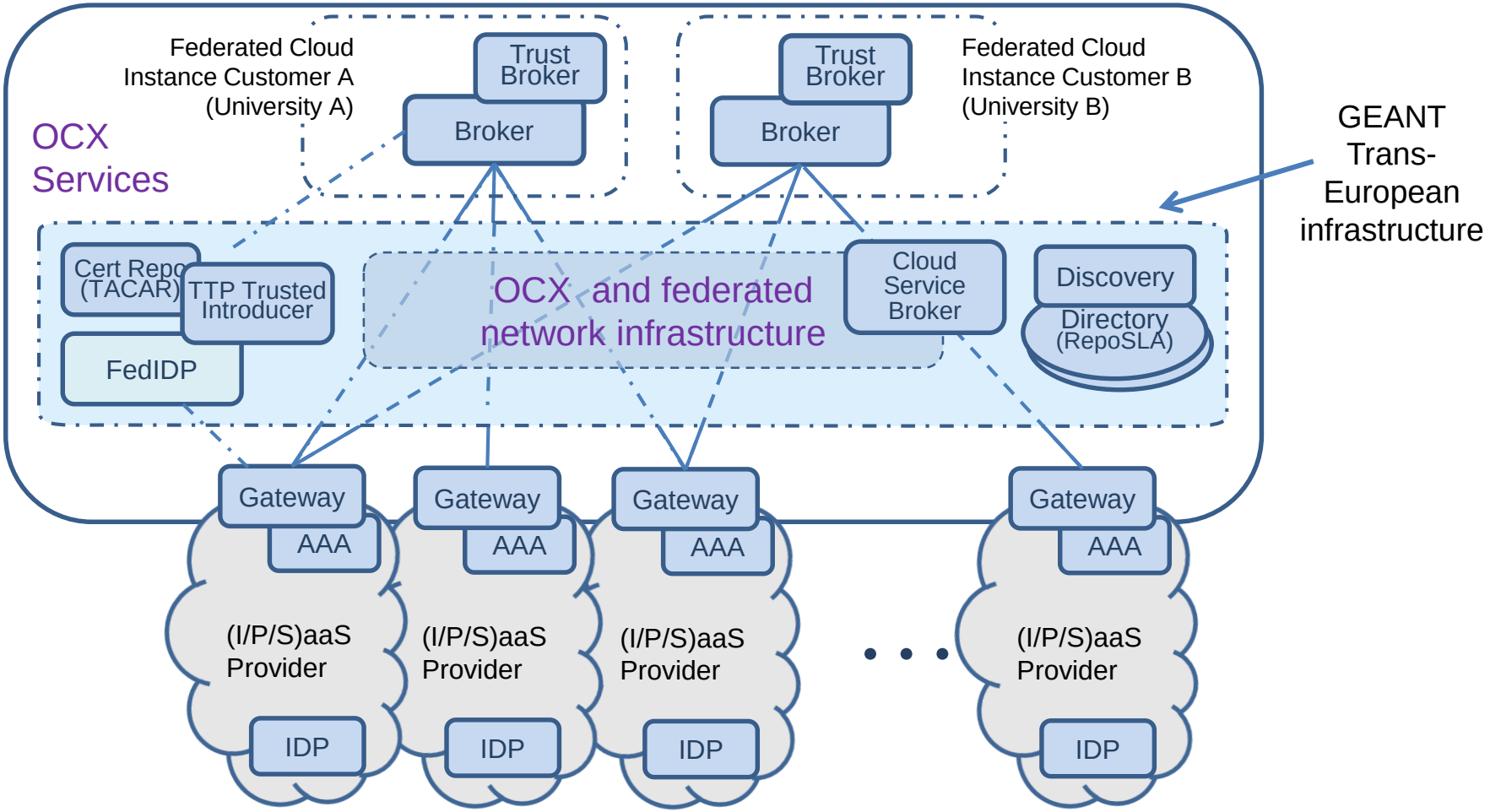
Intercloud Federation Infrastructure and Open Cloud eXchange (OCX)



Provisioning network infrastructure may involve multiple providers: Introducing OCX (Open Cloud eXchange)

OCX (Open Cloud eXchange) Similar to Amazon Direct Connect

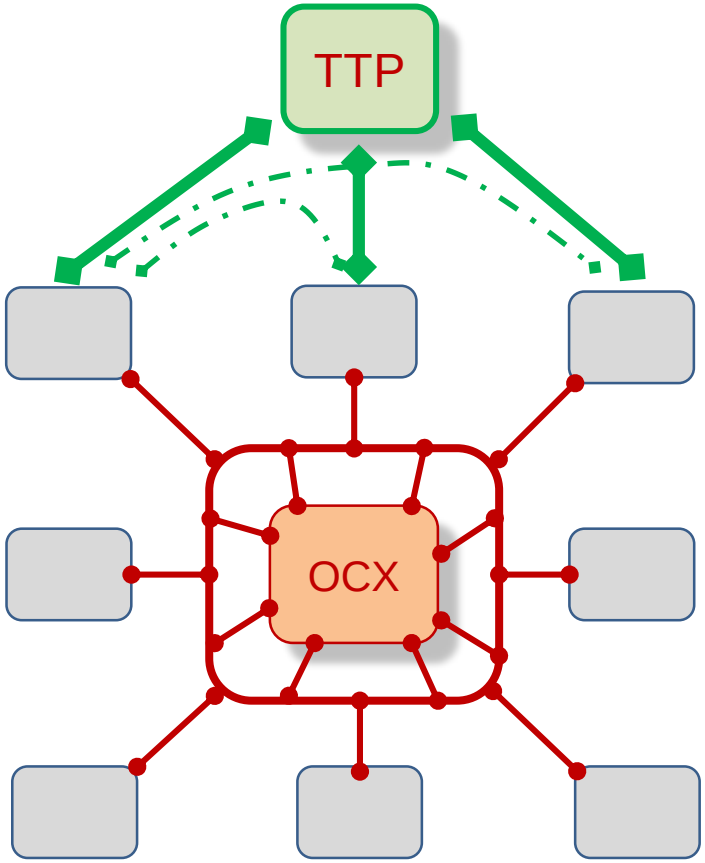
Implementation: Intercloud Federation Infrastructure and Open Cloud eXchange (OCX) in GEANT infrastructure



OCX Definition and Operational Principles

- ***Direct service/inter-member peering***
  - Re-use and leverage Internet eXchange Point (IXP) experience
  - Open collocation services
- ***No third party (intermediary/broker) services***
  - Transparency for cloud based services
  - No involvement into peering or mutual business relations
- ***Trusted Third Party (TTP)***
  - To support dynamic service agreements and/or federation establishment
  - Enables creating federations on-demand
  - *Trusted Introducer for dynamic trust establishment*
- May include other special services to support smooth services delivery and integration between CSP and Customer
  - E.g., Local policies, service registry and discovery, Application/VM repository

OCX Trusted Third Party services



Pre-established trust relation with OCX as TTP

Trust relations established as a part of dynamic federation between OCX members

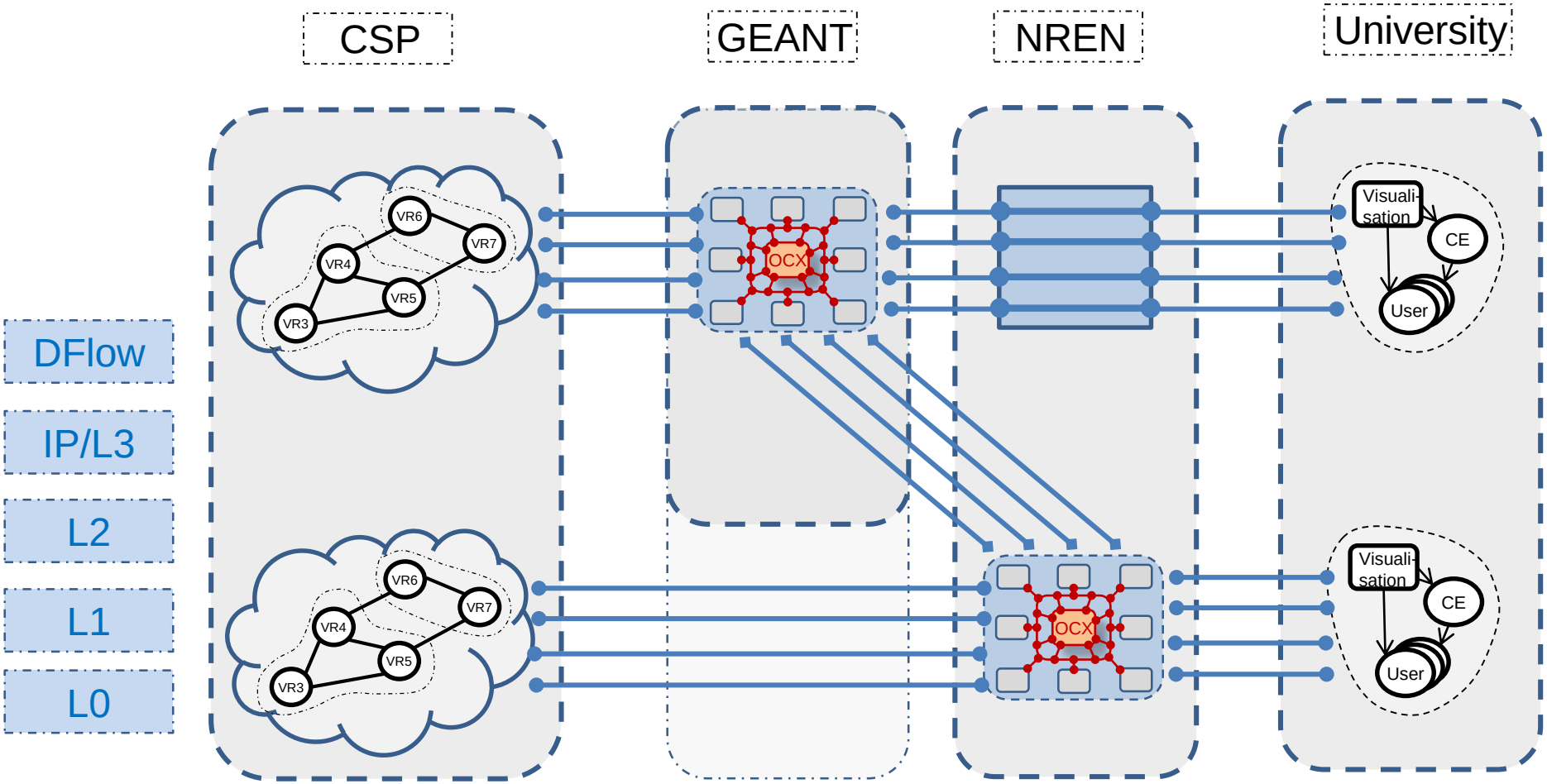
OCX L0-L2/L3 topology

- Any-to-any
- Distributed, collapsed, hierarchical
- Topology information exchange L0-L2 + L3?
- QoS control
- SDN control over OCX switching

TTP goals and services

- Enable dynamic federations establishment
- Trusted Certificates and CA's Repository
  - Similar to TACAR (TERENA Academic CA Repository)
- Trusted Introducer Service
  - Trusted Introduction Protocol
- Service Registry and Discovery
- SLA repository and clearinghouse

OCX Hierarchical Topology Model



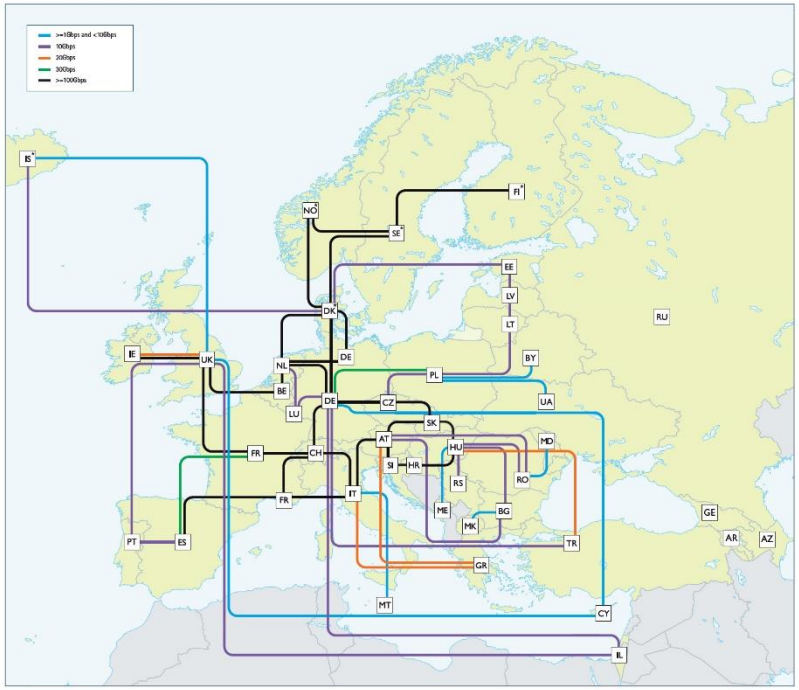
GEANT: European and Worldwide Scale of Infrastructure (2013-2014)

GÉANT

www.geant.net

The Pan-European Research and Education Network

GÉANT interconnects Europe's National Research and Education Networks (NRENs). Together we connect over 50 million users at 10,000 institutions across Europe.



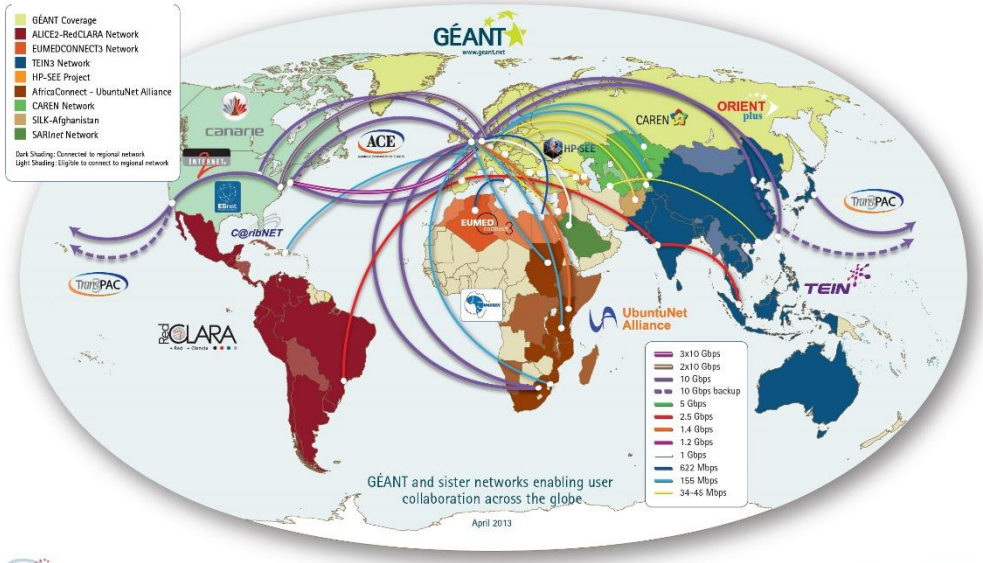
GÉANT connectivity as at January 2014. GÉANT is operated by DANTE on behalf of Europe's NRENs.

DANTE www.dante.net

GÉANT is co-funded by the European Union within its 7th RE&D Framework Programme.

This document has been produced with the financial assistance of the European Union. The contents of this document are the sole responsibility of DANTE and can under no circumstances be regarded as reflecting the position of the European Union.

GÉANT At the Heart of Global Research Networking



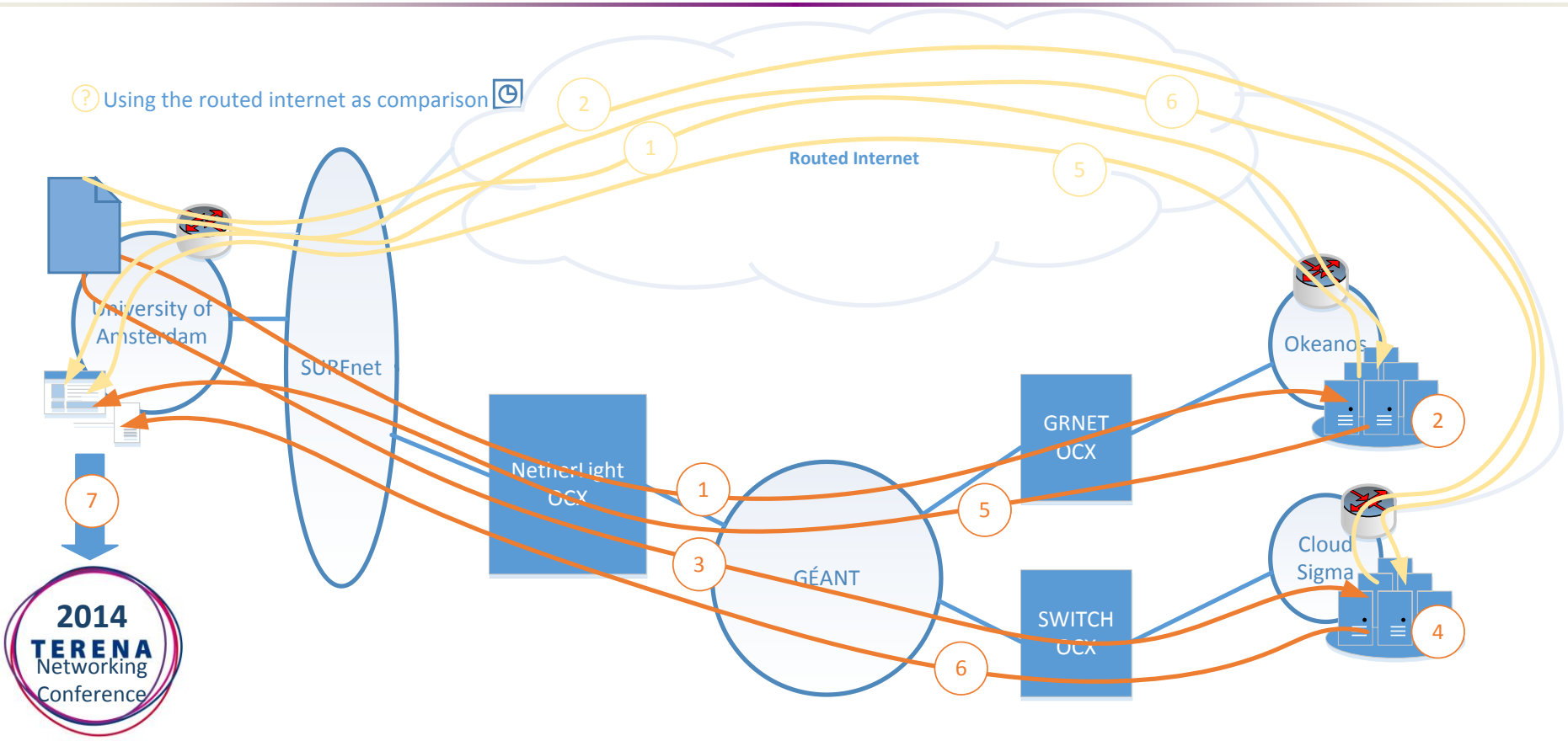
DANTE www.dante.net GÉANT www.geant.net

connect • communicate • collaborate

GÉANT is co-funded by the European Union within its 7th RE&D Framework Programme.

This document has been produced with the financial assistance of the European Union. The contents of this document are the sole responsibility of DANTE and can under no circumstances be regarded as reflecting the position of the European Union.

OCX Pilot: Demo at TNC2014 Conference (19-22 May 2014, Dublin)



Video Processing Sequence

1 - Spawn VMs at Okeanos and send video frames towards these VMs

2 - Transcoding at Okeanos VMs

3 - More CPU power required; spawn VMs at Cloud Sigma and send video frames towards these VMs

4 - Transcoding at Cloud Sigma VMs

5 - Okeanos VMs send transcoded frames to UvA

6 - Cloud Sigma VMs send transcoded Frames to UvA

7 - Show results at TNC

TNC2104 Demo Scenario: HD video editing and streaming

The University of Amsterdam (UvA) has some 4K movies that need efficient transcoding

- Using local OCX (NetherLight) the UvA can get access to necessary compute resources at different Cloud Service Providers via high performance dedicated network links.
  - The demo uses Okeanos (connected via GRNET OCX) and Cloud Sigma (connected via SWITCH OCX).
- The UvA created scheduling software that is able to spawn virtual machines at Okeanos or Cloud Sigma
- The machines are spawned inside the L2-domain of the UvA

OCX enabled GEANT infrastructure provides the following benefits

- Allow the R&E community to select from a broad range of cloud services that ensure network service levels and/or have a logical separation from the Internet
- Allow CSPs to deliver their services efficient, using optimized paths, to the R&E community (everyone is welcome, no limitations on “cross-connects”)
- Facilitate transparent connectivity between the R&E community and CSPs (allow jumbo frames, no firewalls/policies, private network, etc)
- Enhance “time-to-market” by using Bandwidth-on-Demand or other Software Defined Networking (SDN) solutions

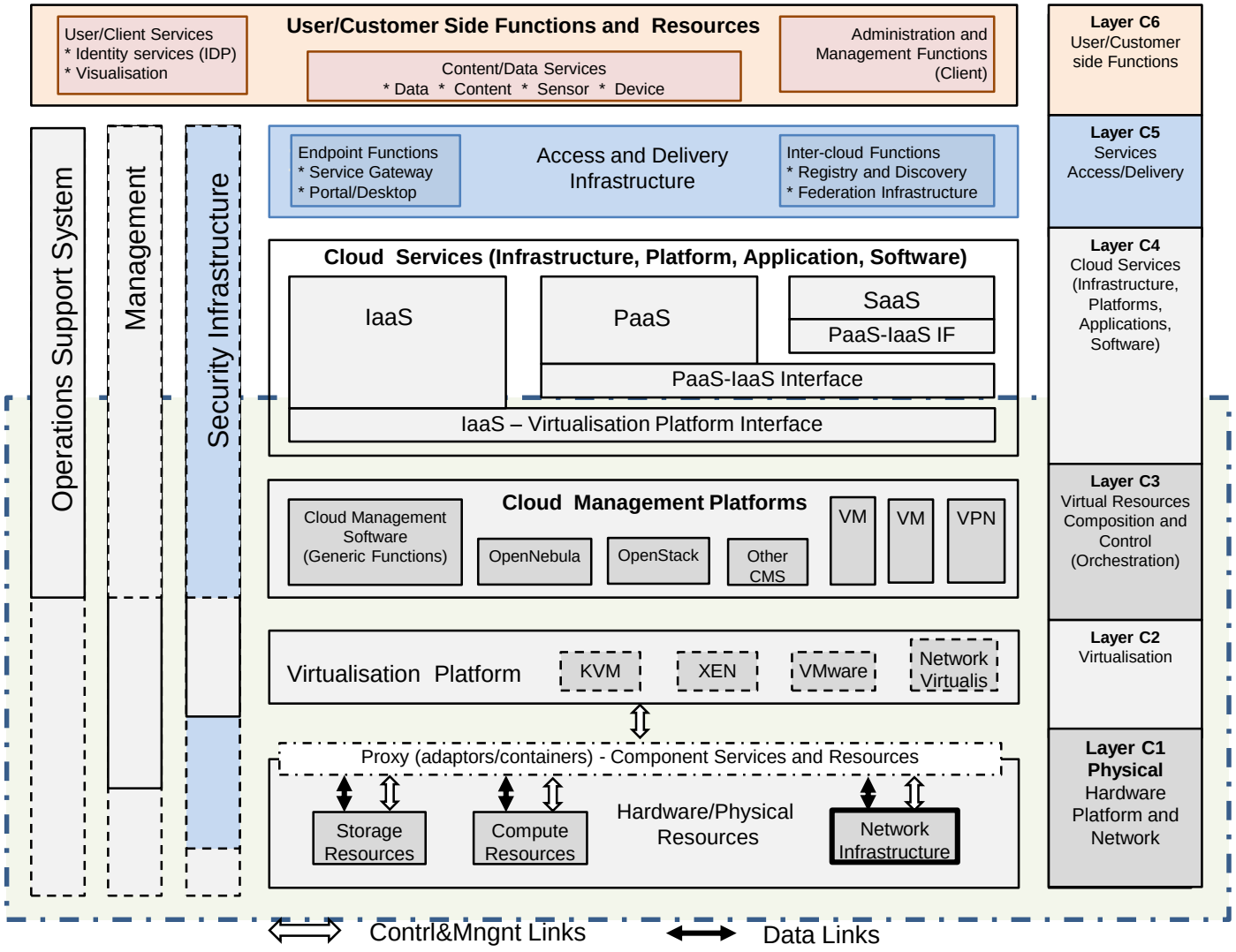
Questions and discussion

- Which cloud federation model to use?
- What research community cloud to join?
- Research grants by the major cloud providers Amazon AWS, Microsoft Azure, IBM

Additional Information

- Cloud Security Challenges and models

Multilayer Cloud Services Model (CSM)



CSM layers

(C6) User/Customer side Functions

(C5) Intercloud Access and Delivery Infrastructure

(C4) Cloud Services (Infrastructure, Platform, Applications)

(C3) Virtual Resources Composition and Orchestration

(C2) Virtualisation Layer

(C1) Hardware platform and dedicated network infrastructure

Control/ Mngnt Links

Data Links

Cloud Computing Security – Challenges

- Fundamental security challenges and main user concerns in Clouds
  - Data security: Where are my data? Are they protected? What control has Cloud provider over data security and location?
  - Identity management and access control: Who has access to my data?
- Two main tasks in making Cloud secure and trustworthy
  - Secure operation of Cloud (provider) infrastructure
  - User controlled access control (security) infrastructure
    - Provide sufficient amount of security controls for user
- Cloud security infrastructure should provide a framework for dynamically provisioned Cloud security services and infrastructure

Current Cloud Security Model

- SLA and Provider based security model
  - SLA between provider and user defines the provider responsibility and guarantees
    - Data protection is attributed to user responsibility
    - Actually no provider responsibility on user run applications or stored data
  - Providers undergo certification of their Cloud infrastructure (insufficient for highly distributed and virtualised environment)
  - Customer/User must trust Provider
- Using VPN and SSH keys generated for user infrastructure/VMs
  - Works for single Cloud provider
  - Inherited key management problems
- Not scalable
- Not easy integration with legacy user/customer infrastructure and physical resources
- Simple access control, however can be installed by user using SSO to Cloud provider site
- Trade-off between simplicity and manageability

Cloud Environment and Problems to be addressed

- Virtualised services
- On-demand/dynamic provisioning
- Multi-tenant/multi-user
- Multi-domain
- Uncontrolled execution and data storage environment
  - Data protection
    - Trusted Computing Platform Architecture (TCPA)
    - Promising homomorphic/elastic encryption (to be researched)
- *Integration with customer legacy security services/infrastructure*
  - *Campus/office local network/accounts*
- *Integration with the providers business workflow*

Emerging Cloud Security Models

- Former (legacy): Provider - User/Customer
- New Cloud oriented security provisioning models
  - Provider - Customer - User
    - Enterprise as a Customer, and employees as Users
    - Enterprise/campus infrastructure and legacy services
  - Provider – Operator (Broker) - Customer – User
    - Application area IT/telecom company serves as an Operator for application services infrastructure created for customer company
- Security issues/problems in new security provisioning models
  - Integration of the customer and provider security services
  - Identity Management and Single Sign On (SSO)
    - Identity provisioning for dynamically created Cloud based infrastructure or applications