PRACE iRODS Workshop

iRODS experience in EUDAT

Giuseppe Fiameni (g.fiameni@cineca.it),
Giacomo Mariani (g.mariani@cineca.it),
SuperComputing, Application and Innovation – CINECA

Linkopig September 2012, Sweden

EUDAT

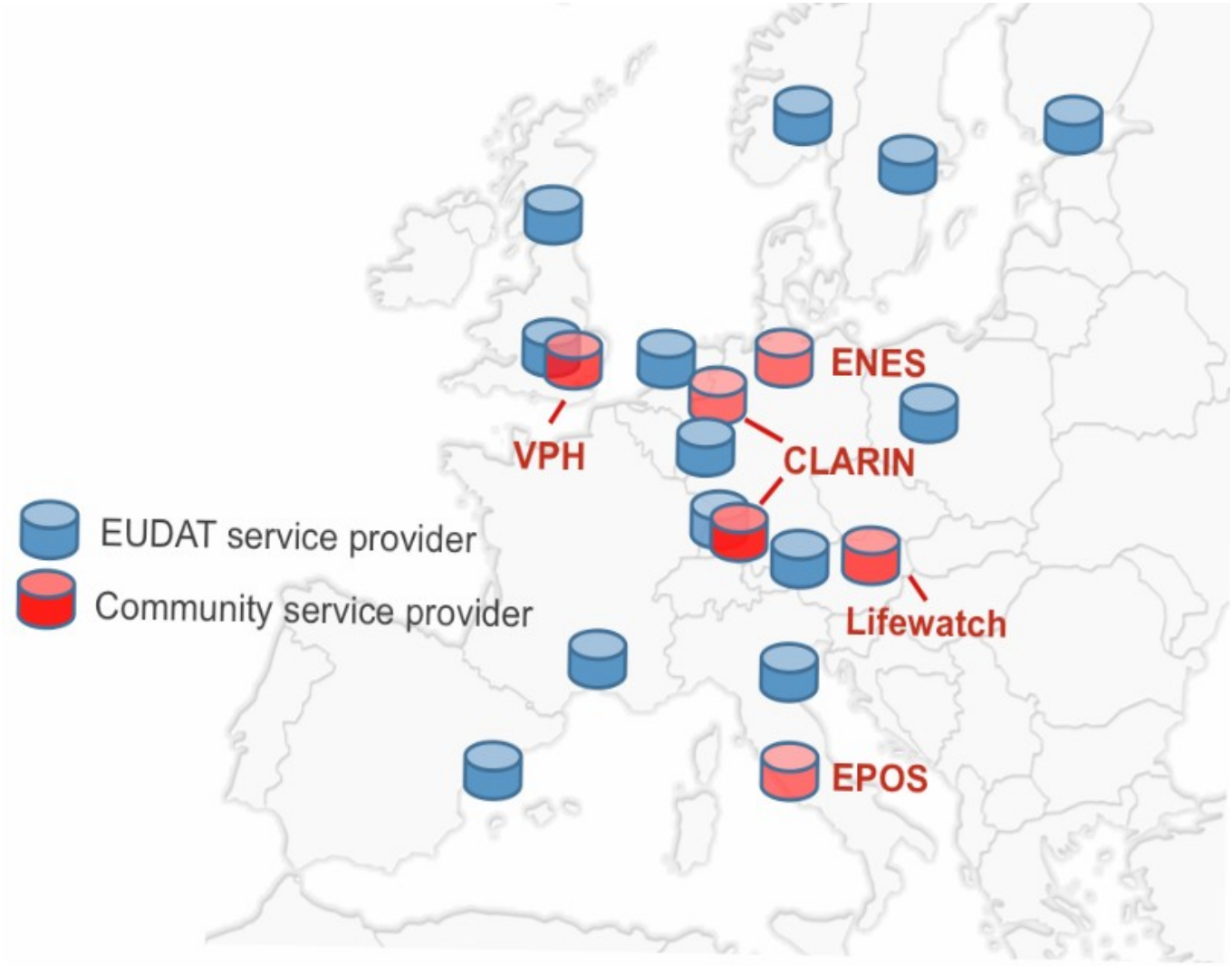
Agenda

- EUDAT
  - EUDAT members
  - EUDAT core services
  - EUDAT approach
- EUDAT and iRODS
  - Safe replication
  - Data Staging
- EUDAT and PRACE

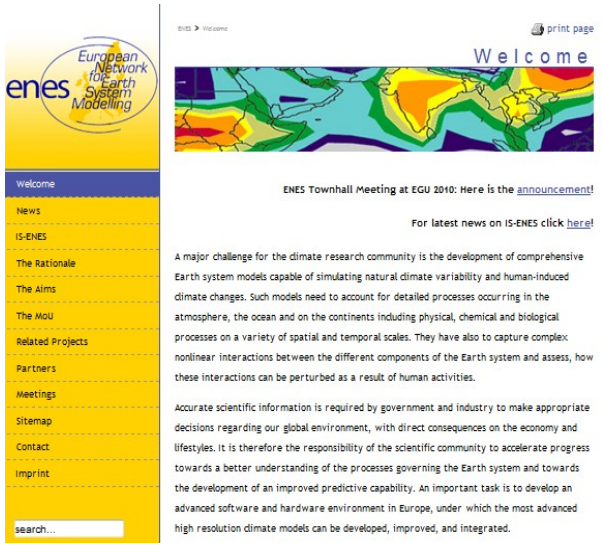
EUDAT: European Data Infrastructure

Project Name	EUDAT – European Data
Start date	1st October 2011
Duration	36 months
Budget	16,3 M€ (including 9,3 M€ from the EC)
EC call	Call 9 (INFRA-2011-1.2.2): Data infrastructure for e-Science (11.2010)
Participants	25 partners from 13 countries (national data enters, technology providers, research communities and funding agencies)
Objectives	“To deliver cost-efficient and high quality Collaborative Data Infrastructure (CDI) with the capacity and capability for meeting researchers’ needs in a flexible and sustainable way, across geographical and disciplinary boundaries.”

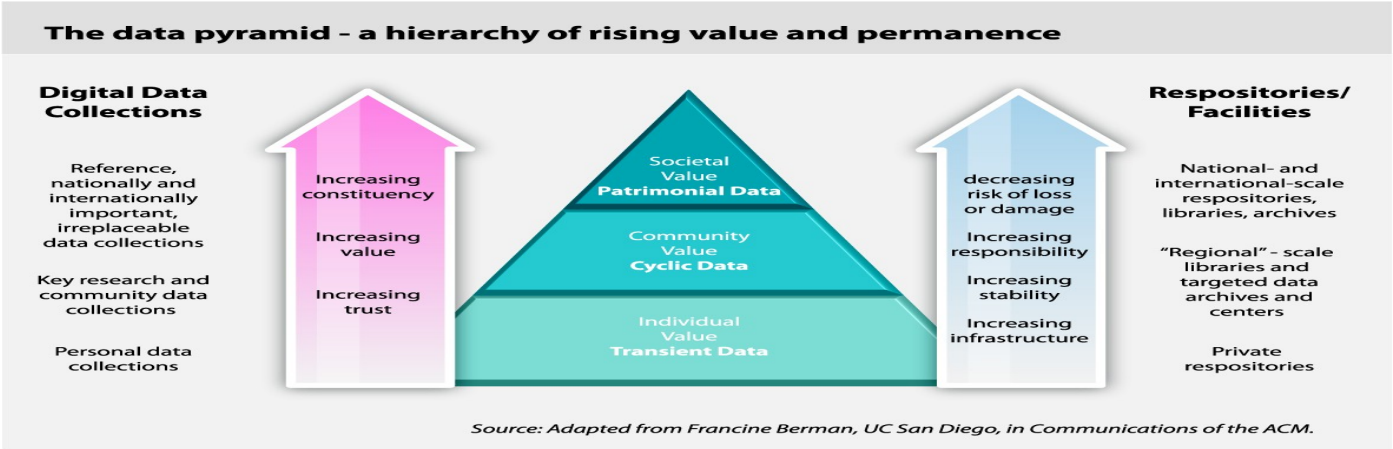
EUDAT members

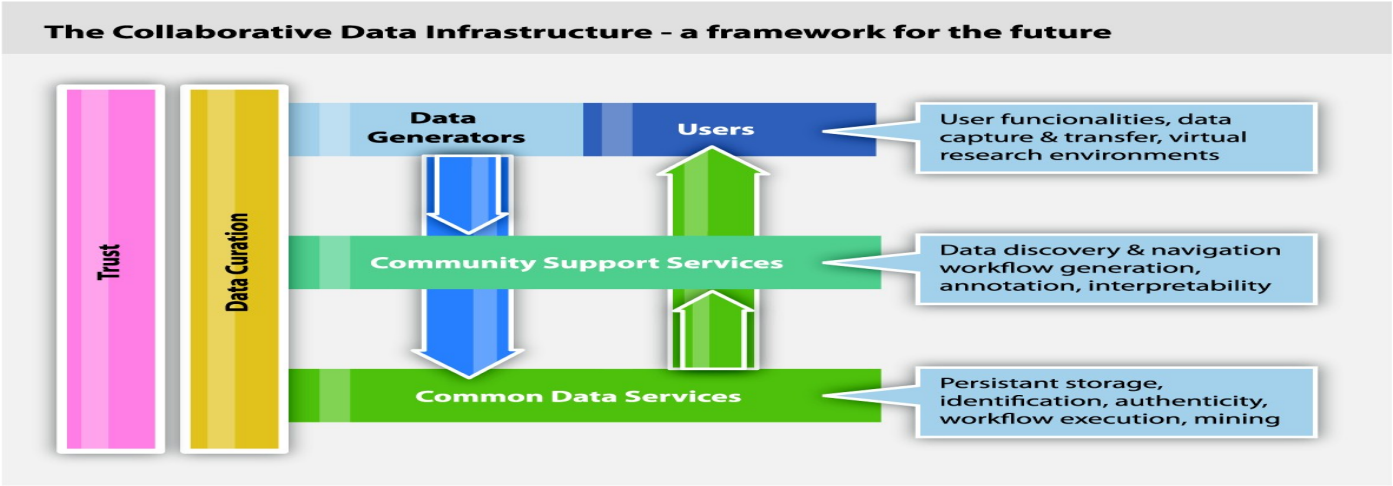


Communities



Common Data Infrastructure





EUDAT Core Service Areas

Community-oriented services

- Simple Data Acces and upload
- Long term preservation
- Shared workspaces
- Execution and workflow (data mining, etc.)
- Joint metadata and data visibility

Enabling services (making use of existing services where possible)

- Persistent identifier service (EPIC, DataCite)
- Federated AAI service
- Network Services
- Monitoring and accounting

Core services are building blocks of EUDAT's Common Data Infrastructure

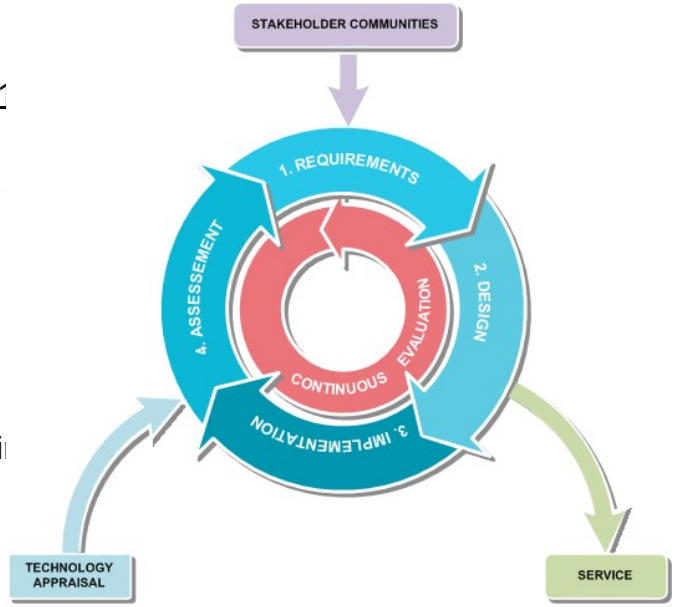
They are mainly included on bottom layer of data services

How to achieve this?

EUDAT service design activities

1. Capturing Communities Requirements (WP4)

- 1st round of interviews with the five initial communities (Oct.-Dec. 201
  - Understand how data is organised in each community
  - Collect first wishes and specific requirements from a common data layer
- Next phase: refine analysis and expanding it to other communities



2. Building the corresponding services (WP5)

- Technology appraisal (ongoing)
  - What is already available at partners's sites to build the correspondi
  - What are the gaps and market failures that should be addressed by
- Next phase: Developing candidate services
  - Adapt services to match the requirements
  - Integrate with community and SP services
  - Test and evaluate with communities

3. Deploying the services and operating the federated infrastructure (WP6)

- Designing the federated infrastructure and the interfaces for cross-site operations (ongoing)
- Next phase: integrating and coordinating resource provision, operations and support

Community Service Wishes

- **In Progress (High priority)**
- Safe Replication of data (for Bit-stream Preservation & Access Optimization)
- Common Authentication/Authorization Infrastructure
- Staging of data onto HPC resources

- **In Progress (Medium priority)**
- Aggregated EUDAT Metadata Domain
- Researcher Data Store (Simple Upload, Share and Access)

- **In Progress as Research Issues (WP7)**
- More elaborate policy rules and federation scalability
- Generic workflow execution framework

Safe Replication 1/3

- Relevant data needs to be replicated from community centers to a number of data centers in a safe way with several purposes in mind:
  - data bit-stream preservation;
  - more optimal data curation;
  - better accessibility of data;
  - identification of data through Persistent Identifiers (PIDs).

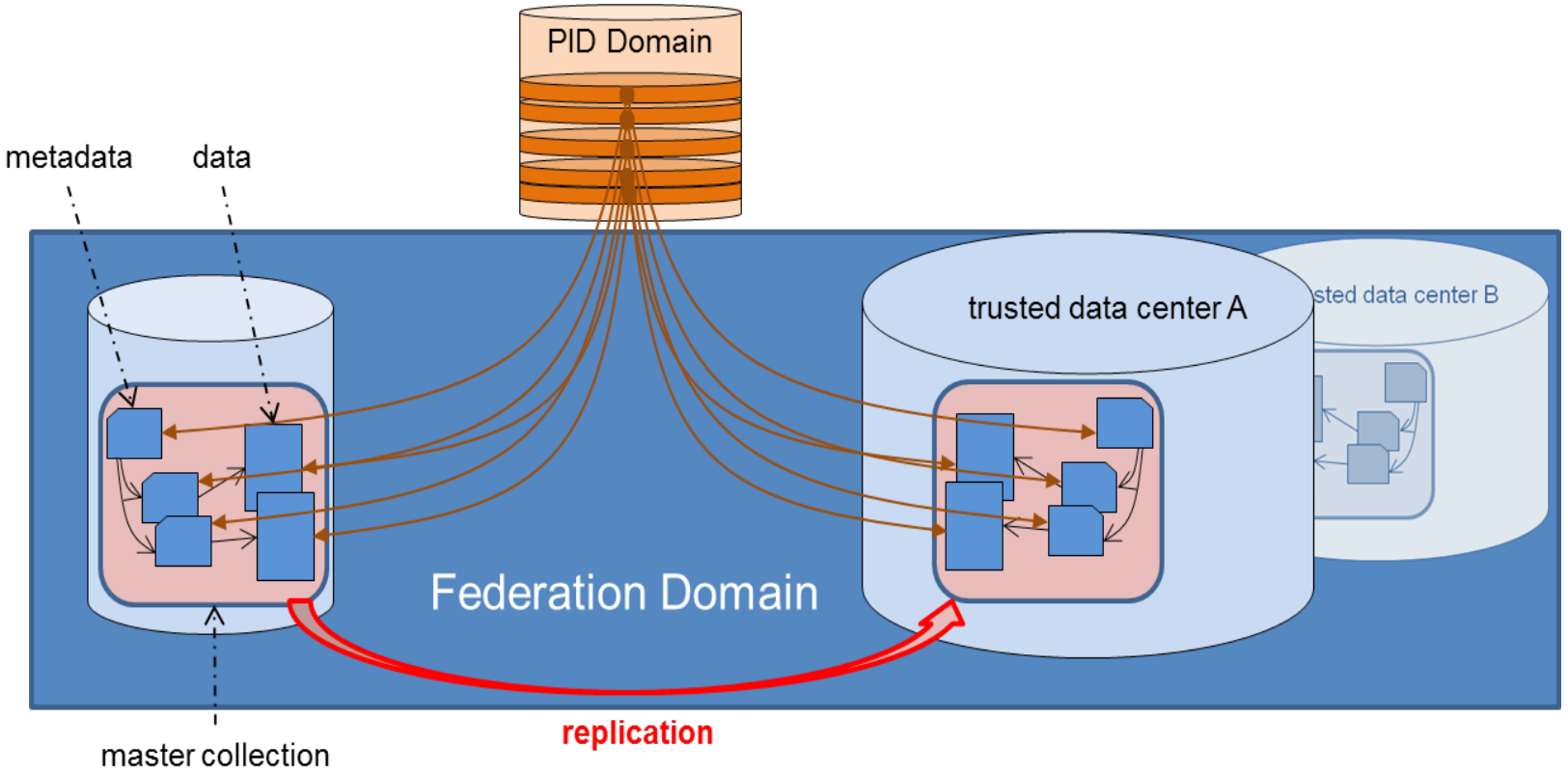
- Common functionality:
  - Create M replicas (identified by a PID record) at different data centers for N years, exclude certain centers, maintaining the given access permissions.

Safe Replication 2/3

- **Technologies:**
  - Long Term Archives → Community specific technologies, through **iRODS UMSS**
  - Policy-based Replication → **iRODS** with rules and microservices
  - Persistent Identifiers → EPIC/Handle

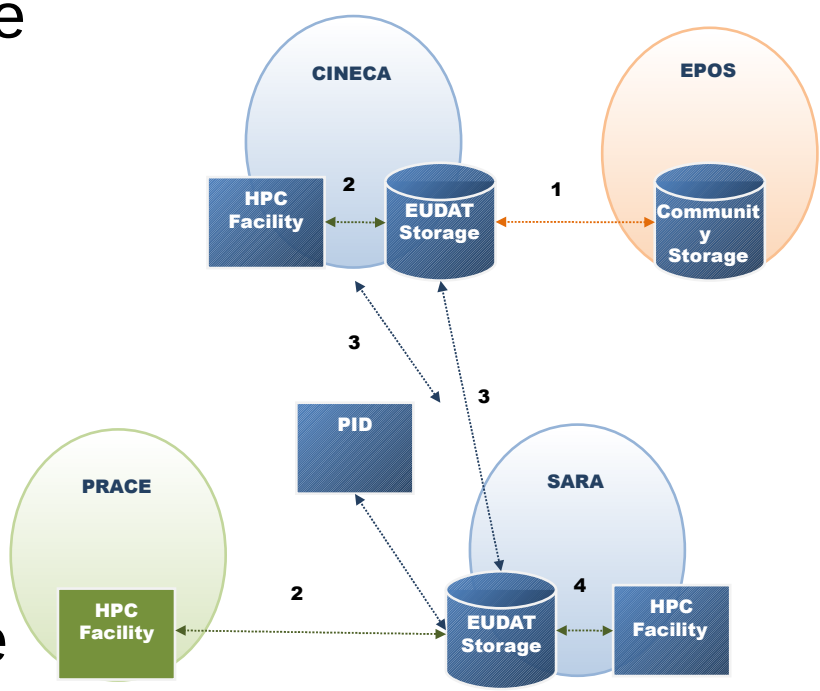
- **Orthogonal aspects:**
  - AAI
  - Monitoring
  - Center Registry
  - Metadata

Safe Replication 3/3



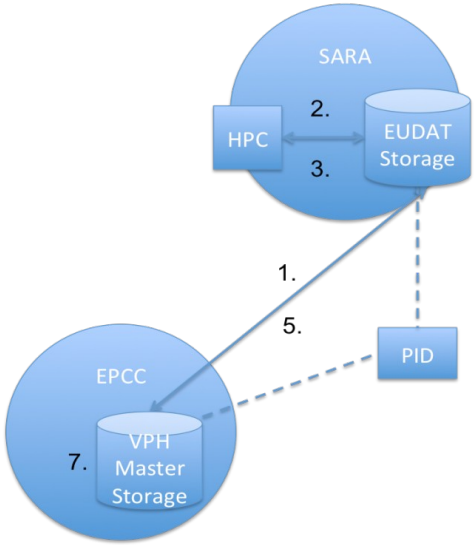
Data Staging 1/3

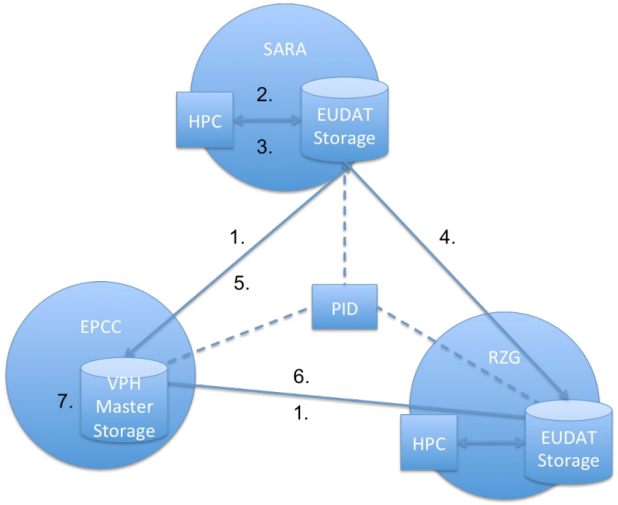
- **Objective**:
  Allow communities to stage data between EUDAT resources and HPC/HTC resources, including the PRACE Infrastructure, for computational purposes.
- **Description**:
  This service will allow the communities to dynamically replicate a subset of their data stored in EUDAT to an HPC machine workspace in order to be processed.



Data Staging 2/3

- Intention is to make use of HPC machines for computations on stored data.
- Different configurations possible:
  - computations on a single HPC node where data already is;
  - computations on multiple nodes - use of PRACE fast distributed file system.
- Principles:
  - user issues a compute command;
  - script pushes data into the HPC workspace, results go into workspace;
  - data, identified by the PID, are retrieved from the “best” resource, transparently;
  - input data is discarded after job





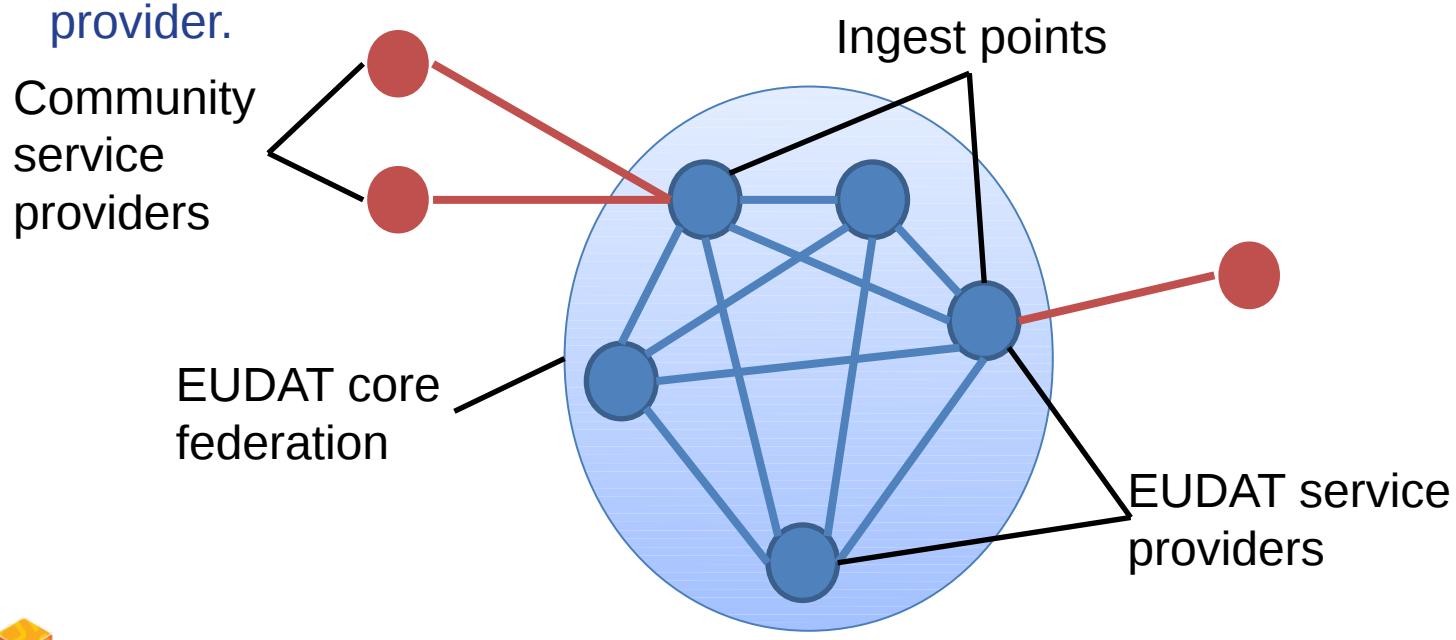
Data Staging 3/3

- **Technologies:**
  - Data transfers -> GridFTP interface to **iRODS (Griffin)**
  - Data access and selection -> various interfaces/clients (XSEDE File manager, Globus Online)

- **Orthogonal aspects:**
  - AAI
  - Monitoring

EUDAT and iRODS, the core technology

- iRODS is installed in all EUDAT service providers and community service providers.
- Data are stored in iRODS Vault or available as mounted collection.
- The iRODS instances are federated following the “snow flake” approach: each community service provider is federated with an ingest point, i.e. an EUDAT service provider federated with all the others EUDAT service provider.



iRODS used functionalities

- iRODS, version 3.0 or 3.1, was tested during the first technology appraisal activity.
- The services will be moved in production with iRODS 3.1 and probably be moved to iRODS 3.2 as soon as it will be available.
- The metadata structure definition is on going, so metadata are not used at the moment.
- embedPython has been used with good results, but, due to a recently discovered security issue, its usage has been suspended.
- Griffin, a JAVA GridFTP server able to exploit iRODS as a resource, is in use.
- Davis, a JAVA WebDAV interface has been tested, but it shows some issues (for example some crashes moving big files).
- iDROP web will be probably considered for the web interface.

TF goals and used iRODS functionalities

	iRODS ACL	iRODS rules and microservices	Embed python	griffin	UMSS
Staging to HPC	Yes			Yes	
Replicate in different zones	Yes	Yes	Optional		
Share data	Yes	Yes			
LTA					Yes
PID		Yes	Optional		

CINECA experience

- iRODS 3.0 installed with support for GSI, FUSE and OS authentication
- External mySQL farm
- Data stored in a Vault exploiting GPFS over NFS
- Compound resource exploiting TSM via UMSS
- Ingestion of more than 5 TB to TSM
- Daily sync of about 5TB with a federated zone
- Installed and tested Griffin (solved bugs with GO)
- Installed and tested Davis
- Developed a script for Data Staging exploiting the GO API
- *Performance issues with TSM and real time*

EUDAT experience

- Tens of iRODS installations and federations
- Evaluation of irods configurations
- Vault and imcoll deeply investigated
- Various federations scenarios evaluated
- Various LTA technologies connected
- Performance evaluations moving data and synchronizing data
- Experiences transferring TB of data (big files as well as losf)
- Rules and microservices for replication and PIDding developed
- Griffin evaluation
- Run as root in order to preserve ownership considered

Technical details at CINECA

- Debian 6.0 on a 8 processors (x86_64)
- iRODS 3.0
- Python 2.6 (2.7 from pythonbrew)
- MySQL 5.1
- Perl 5.10.1
- Griffin 0.8.5
- Davis 0.9.4

EUDAT wishes

- LDAP integration
- Packaging for easier updates
- Optional use of system packages (such as system DB or GSI libraries)
- Migration tool able to import an archive already existing on a shared FS

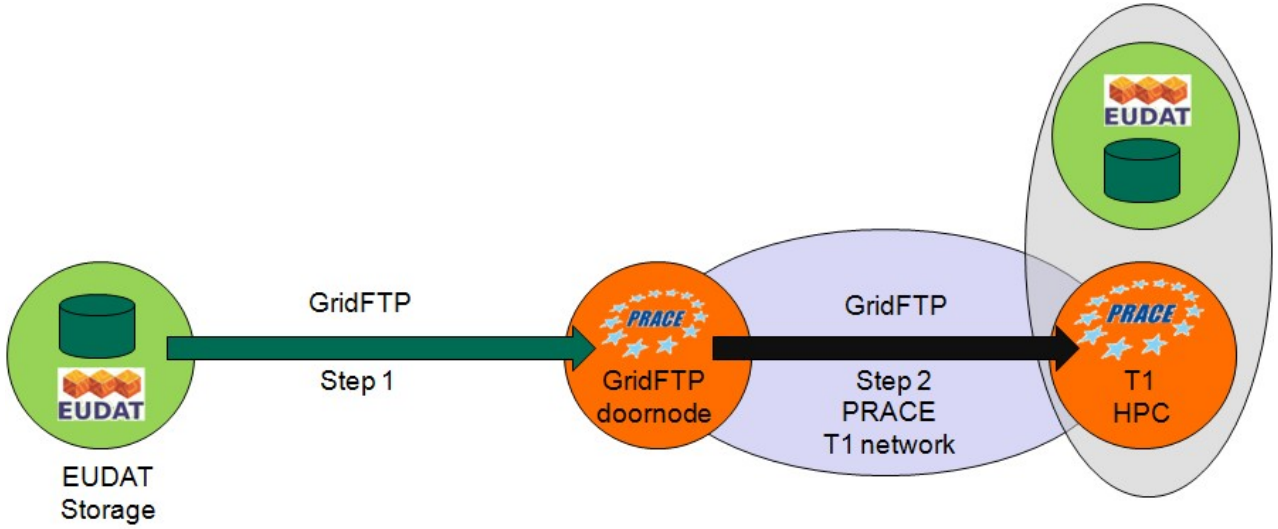
Other, small whishes

- Fixing the awkward behavior of remote users in rules execution: for example, a user defined in zona A cannot invoke rules in zone B where a irsync from zone B to zone A is involved.
- Fixing third party transfer: a user defined in A and defined as remote in zone B and zone C cannot move data from B to C.

Collaboration with PRACE (background)

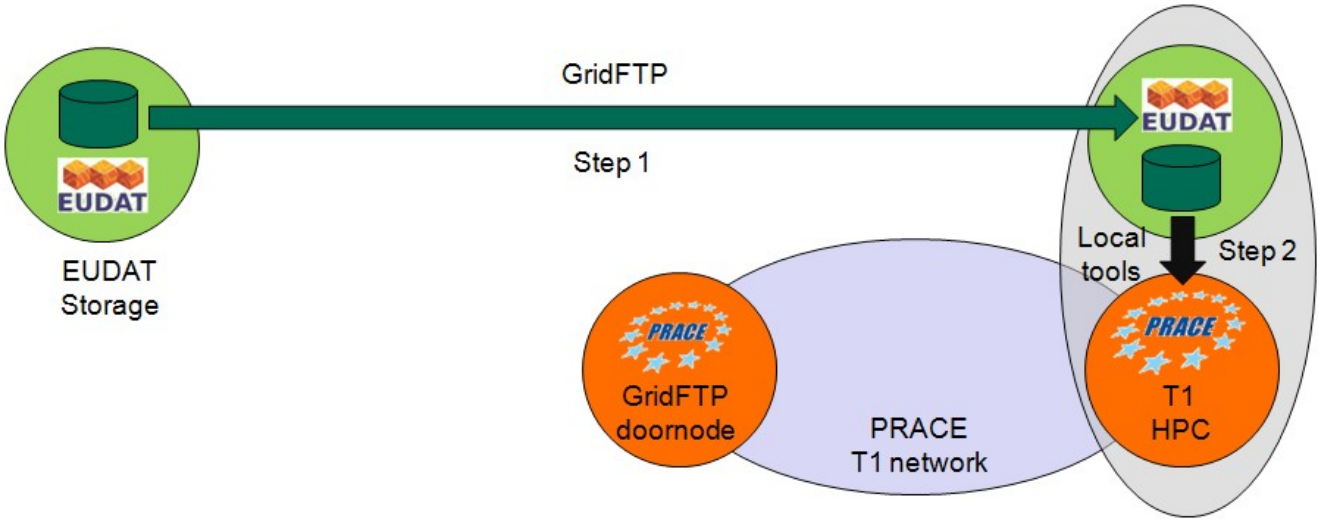
- Some EUDAT Communities (EPOS, VPH) manifested their interest in staging data onto and from PRACE resources.
- Many EUDAT sites are member of PRACE either.
- Staging of data within the PRACE Infrastructure is enabled through the GridFTP protocol under precise conditions: a) staging of data from the public network is permitted only through PRACE door nodes, namely CINECA and LRZ; b) staging of data within the private network is permitted without any restriction through

Two possible scenarios (1)



- Data transit through the PRACE door node before reaching the target HPC facility

Two possible scenarios (2)

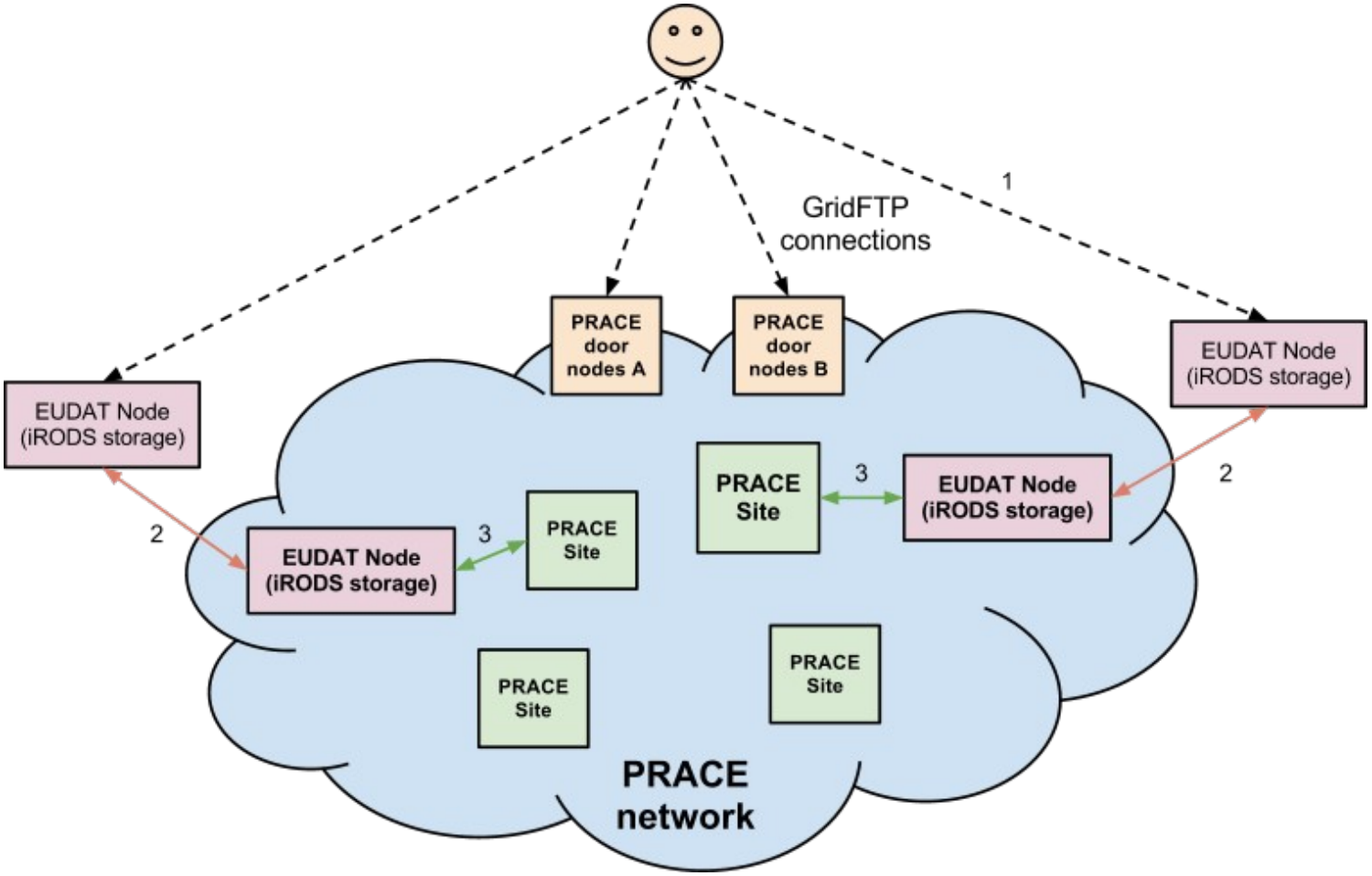


- Data a replicated onto an EUDAT node which is “closer” to the target HPC facility. What happens to those sites being not part of PRACE?

Possible collaboration plan (1/2)

- Few EUDAT nodes enter the PRACE internal network. Data to be staged on PRACE facilities are first replicated to one of these nodes and then staged to the final destination. This scenario will open a new collaboration perspective as, on one hand, EUDAT/PRACE users would be granted with the possibility to stage their data to PRACE facilities, on the other, PRACE users could be provided with the possibility to deposit their results into EUDAT storage and request EUDAT to preserve them. *An ideal scenario would be that PRACE users, during the submission of their*

Possible collaboration plan (2/2)



Thanks for your attention

- Q&A
- :wq