



# C3-INAD and ESGF: Integration strategy

## C3-INAD Middleware Team:

Stephan Kindermann, Carsten Ehbrecht [DKRZ]

Bernadette Fritsch [AWI]

Maik Jorra, Florian Schintke, Stefan Plantikov [ZUSE Institute]

Markus Kemmerling, Christian Grimme [TU-DO]



# C3-INAD Overview

- C3Grid-INAD (08.2010 – 07.2013) (Climate community specific funding)
- follow up of C3-Grid (09.2005 – 03.2009) (D-Grid funding)

## Key characteristics of C3-Grid:

- National Climate Data Center Integration (world data centers, research institutes, universities)
- ISO 19139 Metadata, uniform data access interface, Portal
- Grid based distributed workspace with data/job co-scheduling middleware

## C3-INAD = C3-Grid +

- Production level, sustainability plan, support center, new scientific workflows
- ESGF data node integration



# C3-INAD and ESGF

- **Objective:** establish C3-INAD as an alternative to „download and process at home“ approach for ESGF data node hosted data
- C3-INAD develops CMIP5 multi-model-multi-ensemble WFs, e.g. for means, variances, percentiles, statistical means (different ‚multi-model‘ weighting tools) as well as model verification/validation WFs.
- Redesign of C3-Grid infrastructure necessary (no globus WSRF, towards REST architecture) including security architecture (-> opportunity to align developments with ESGF)



# Overview

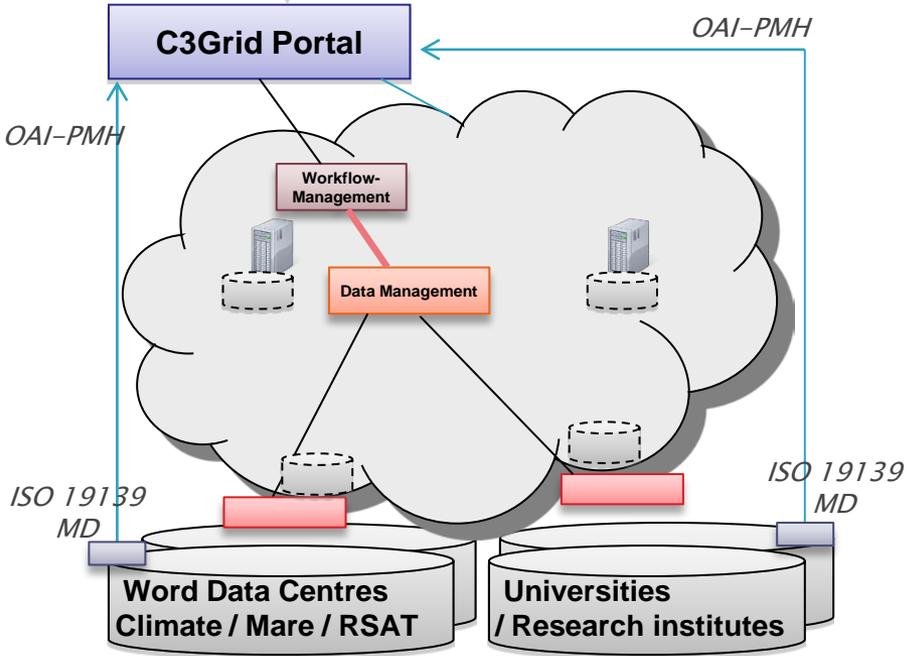
## C3-Grid / C3-INAD infrastructure

- Status / generic architecture
- Data management / job management infrastructure

## C3-Grid / ESGF integration

- Data management / (Security)
- Metadata infrastructure

# C3-Grid: Architecture



- Data Providers deploy a uniform data access interface, which is part of a distributed data management middleware
- Data Providers provide ISO 19139 data discovery metadata, via OAI-PMH
- The portal submits JSDL work flow descriptions to the workflow management system. A workflow consists of data staging and computing tasks
- The data management infrastructure cares for data staging, data replication, data life-time etc.
- Data management and job management components interact via a co-scheduling protocol

Security: X509 (proxy) certificates + (GT4) delegation



# C3-Grid Data Management Infrastructure



„Generation N Data Management System (GNDMS):

<http://gndms.zib.de>

## Key characteristics:

- Based on data staging and co-scheduling
- Data integration layer (logical names, data transfers, workspace management)
- Support for GT4 – transition to REST based architecture

## Features:

- **Co-scheduling** support for all data management activities
- **Failure recovery** for all data management activities
  - persistent storage of whole critical data management status
- **Highly configurable and extensible**
  - generic data management tasks
  - Different Roles
    - C3Grid DMS      Overall C3Grid DMS coordinator
    - Provider      Provide data in C3Grid via staging
- System management components: **Runtime configuration and monitoring**
- Uses **Globus/WSRF, GSI** and **GridFTP** /  
New **REST** based implementation ongoing

# GNDMS Components



- **DSpace**

- data space / workspace manager
- “slice” per job, lifetime, auto delete

- **GORFX**

- generic offer request factory
- offers services depending on site role

- **Available Services:**

- **SliceStageln**

staging at DMS: provider selection and trigger  
remote ProviderStageln

- **ProviderStageln**

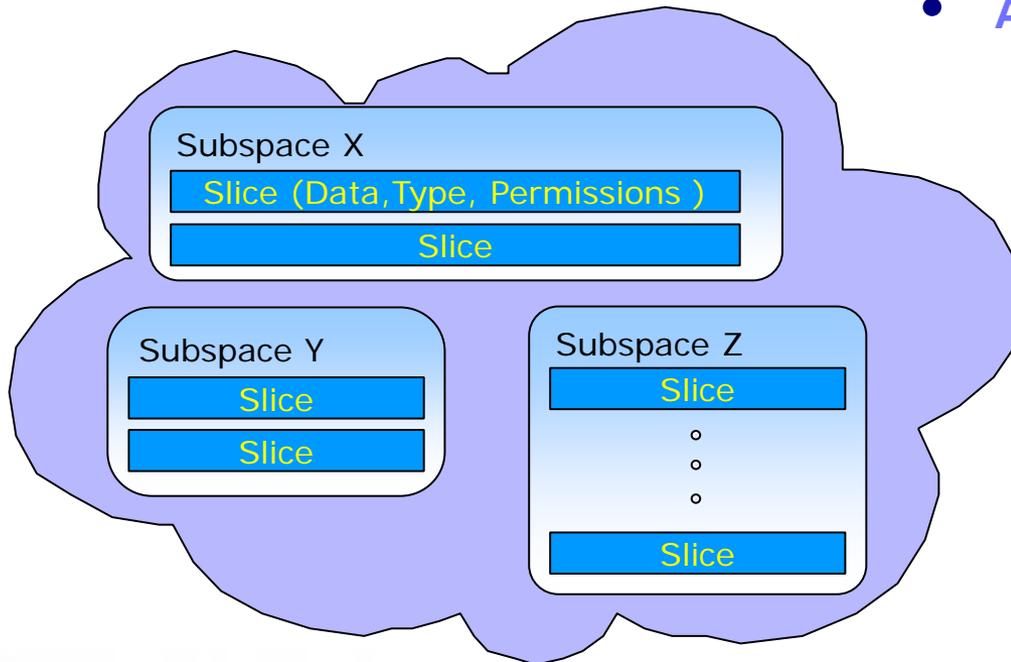
staging into slice at data provider

- **InterSliceTransfer**

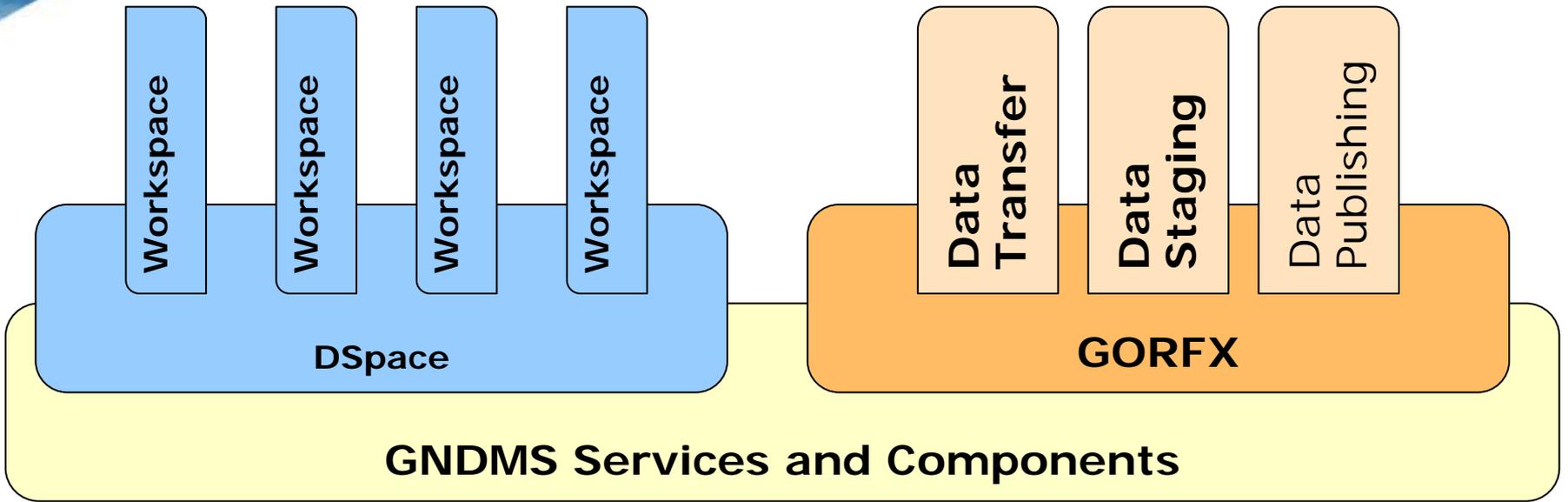
copying of files between slices

- **FileTransfer**

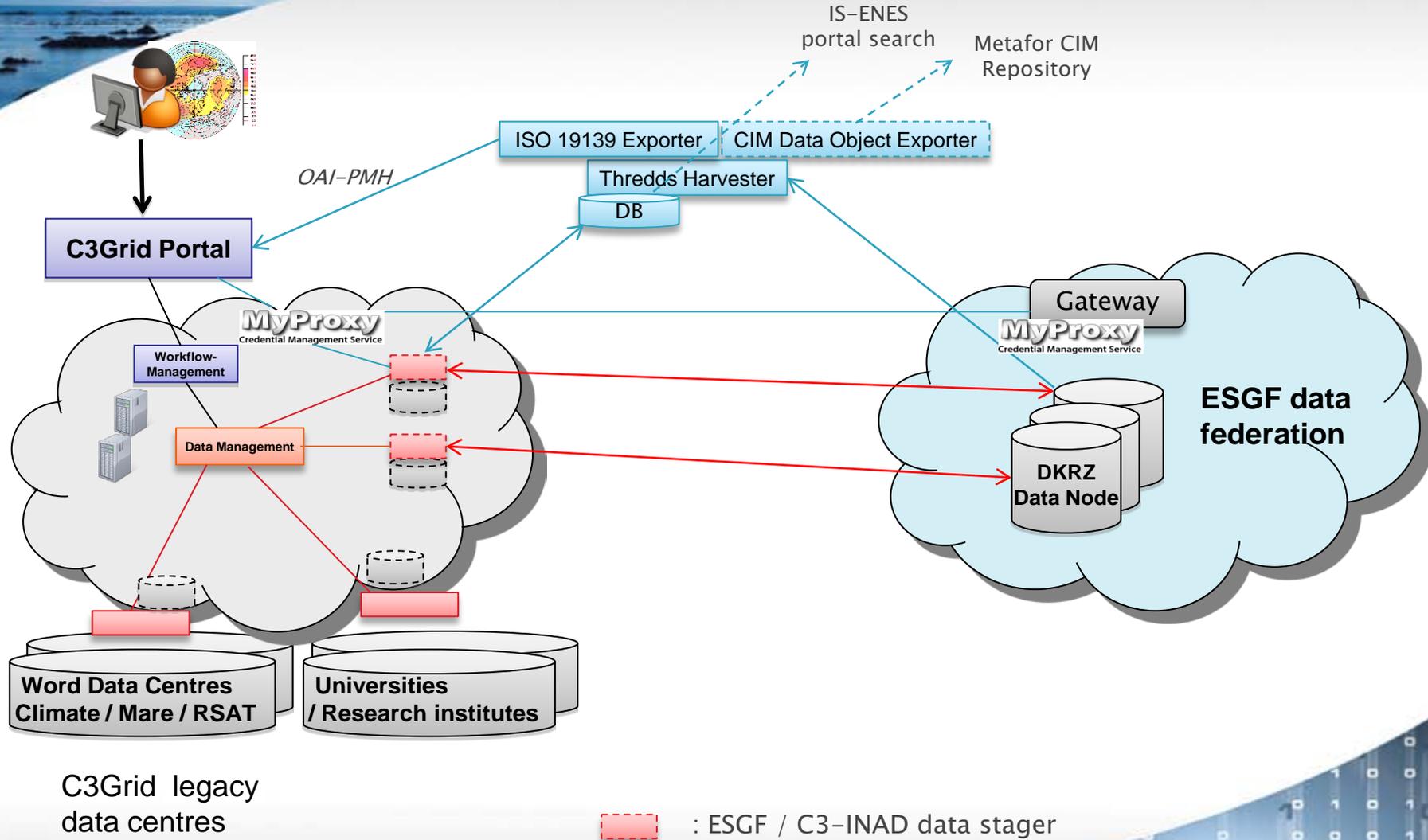
import/export files via GridFTP



GNDMS negotiates *contracts* with clients about task execution. A contract specifies what is to be done, and optionally when and where it is to be done by GNDMS on behalf of the client. *Contracts* are accepted *Offers*



# C3-ESGF integration





# C3 – ESGF integration

## Metadata:

- Dataset metadata available in central DB (Thredds harvest sink + CERA)
- Thredds to ISO 19139 MD generation → harvesting into C3 Portal

## Data: a C3–ESGF data staging request:

- queries central DB for corresponding files and generates access script(s)
- gets delegated certificate and stages files directly from ESGF data nodes
- Sub-selects requested spatio-temporal subset(s), ...
- Generates provenance metadata in ISO 19139 for result

## Security:

- MyProxy based delegation
- O-AUTH is currently being evaluated



# C3 – ESGF next steps

- C3 – ESGF portal integration and data access prototype planed for fall 2011
- Multi-model multi-ensemble wflow deployment in 2012
- Under discussion: ESGF data node sw stack as default way to add new data providers to C3-INAD

Developments and roadmap can be followed at <http://wiki.c3grid.de> (registration necessary)

There you will also find info on reusable components like

- Thredds harvester and DB ingester + ISO generator
- GNDMS middleware
- ...



# Summary

## Old C3 – Grid:

- Data center integration, portal, middleware development

## New C3 – INAD:

- Climate scientist, use case driven (C3-INAD „workflow developers“)
- ESGF data node integration
- new development of middleware components necessary → opportunity to align developments with ESGF