A Collaborative Environment for Climate Data Handling

- C3Grid -

Stephan Kindermann, DKRZ
Martina Stockhause, MPI-M
C3-Team
The mission for C3Grid:
Build up a collaborative environment to ease **data discovery, data access and data processing**

Being part of a national Grid Infrastructure
The C3Grid Consortium

Scientific Users

MPI-M (Max Planck Institute for Meteorology)
IFM-GEOMAR (Leibniz Institute of Marine Sciences)
University of Cologne
Freie Universität Berlin
PIK (Potsdam Institute for Climate Impact Research)
DLR (German Aerospace Center)
AWI (Alfred Wegener Institute for Polar and Marine Research)
GKSS

Associate Partners

University of Hannover
University of Bonn Forschungszentrum Karlsruhe

Further info:
http://www.c3grid.de

kindermann@dkrz.de

Partners in information sciences

University of Dortmund
Zuse Institute Berlin (ZIB)

Scientific Data provider

World Data Centers
WDC Climate
WDC Mare
WDC RSAT

DWD (German Meteorological Service)

DKRZ (German Climate Computing Center)
PIK (Potsdam Institute for Climate Impact Research)
AWI (Alfred Wegener Institute for Polar and Marine Research)
IFM-GEOMAR (Leibniz Institute of Marine Sciences)
University of Cologne/ ZAIK/ FUB

Sun Microsystems, Inc.
NEC Corporation
Brockmann Consult
C3Grid in context

**Climate Community / Organizations**
GO-ESSP, CF, IPCC, Earth System Curator ...

**Other Grids:**
EGEE, ESG, NDG

**Geosciences:**
INSPIRE, ISO 19115, OGC

**New FP7 projects**
Metafor, IS-ENES, Europs-Net ..

**Security:**
Shibboleth, SAML

**Grid / Web Basics:**
OGF, Globus, WSRF, GSI, WSI, ..

D-Grid

**Application**

1. 6. 2008

C3Grid

GeoInformatics
C3Grid Architecture

Key Challenges / Problems

- Data discovery
- Data access
- Data processing
- Access to resources

- Automatic Metadata Generation
- Bridging Heterogeneity
- Workflow Composition
- Consistent Security Infrastructure
Discovery of data products:

Open standards:
- OAI-PMH harvesting
- ISO 19115/19139 Metadata Profile
- freely available discovery framework
  (see www.panfmp.org or sourceforge.net site)

ISO 19139 Profile:
- controlled vocabularies
  (CF standard names, Coordinate Ref. Systems, ...)
- hierarchical data organisation (parent – child, aggregation info, ..)
- data provenance (ISO lineage information)
- data access info (C3Grid specific)
Community Data Access: Bridging Heterogeneity

- **Generic Data Request Web Service Interface**
- **Analysis**
  - Selection of preprocessing tools
  - Metadata generation
- **Provider specific data access interface**

**Primary data**

**Compute Resource**

**Grid based data management**

**Initial Implementation**:
- WSDL Web Service
- next: WSRF Web Service

**Implementation Examples**:
- DB + Archive Wrapper (DKRZ, M&D)
- Data Warehouse (Pangaea)
- OGSA-DAI + DB (DWD)
- ...

**Metadata**

- geographical
- + vertical
- + temporal
- + content
- + file format selection

**Base data**

**workspace**

**pre-prec.**

**metadata**

**netCDF, GRIB, HDF, XML, ..**
The C3Grid SOA

- Portal
- DIS
- DMS
- RIS
- Workflow Scheduler
- primary metadata
- primary data
- base data
- Compute Resource
- pre-proc.
- workspace

Distributed grid infrastructure

C3Grid Data / Compute Providers

World Data Centers
- Climat
- Mare
- RSAT
- DWD
- DKRZ

Research Institutes
- PIK
- GKSS
- AWI
- MPI-M
- IFM-Geomar

Universities
- FU Berlin
- Uni Köln

Local resources and interfaces

12. 6. 2008
Problem: Automatic Metadata Life Cycle Management

Problem A:

**Discovery of dynamically generated/inserted data products**

→ Automatic (discovery) metadata generation as part of workflow, including:
  - automatic provenance info tracking
  - security info
  - aggregation info

→ Automatic quality check

→ Generic tool support
Metadata Handling Support

**Browser Interface:**
- XML content visualisation
- XML (template) generation

**API:**
- workflow integration
- e-science middleware integration

**Technology:**
- XML database (eXist), XForms (Orbeon), python tools

→ First reusable prototype, cooperation with EGEE
• For established data centers ISO discovery metadata generation was quickly done

• Substantial support necessary for generating or manipulating ISO discovery metadata for others

• Agreement on, maintainance of, controlled vocabularies very important

• „USE“ / DISCOVERY metadata gap is an open issue (e.g. GRIB usage problem, ...)

Discovery Metadata: Experiences, Problems
**Problem B:**

Flexibility of processing functionality ??

- predefined workflows (just parameterization)
- workflow composition based on module library
- user inserted programs
Workflow Processing: Status / Next steps

• Initial set of fixed workflows integrated in portal

• Scheduler controls execution (decision based e.g. on modules info + data availability)

• „modules“ info published to Grid Resource Information Service (MDS based)

• GT4 WS-GRAM Interfaces

• Preinstalled SW packages (use of „modules“ system)

Open Issues:

• workflow composition support → interdependency between processing and data

• user defined processing → debugging, substantial user support needed; security!
Data access and security

- single sign on
- support of users without grid certificates
- federated identity management

- X509 grid certificates (EU-GridPMA CA)
- Grid security infrastructure (GSI)

- legacy AA infrastructure (LDAP, DB based, ..)
- legacy data access infrastructure
C3Grid Security Infrastructure

Identity Provider
Home Organisation

Identity Provider
Virtual Organisation

"Home attributes + VO attributes"

Portal

GridShib
SAML tools

wflow client

X509 Grid-proxy

<..SAML Assertions...>

SLCS (CA)
MyProxy

Delegation Service

Grid Service

Grid Resource

SAML

GridShib for GT

GRAM / DataRAM

Personal / Group Account

C3Grid Middleware
Approach:

2. Shibboleth (WEB2.0) federated identity management, Home + VO attrib.

3. X509 Grid certificates + GSI in Grid Infrastructure

1 + 2: SAML assertions → short lived certificate from SLCS CA

Problem:

• *Initial Idea was:* „Take and adapt existing technology“

• *Status is:* ambitious, risky security infrastructure establishment

• new technology, often partly documented

• no working security infrastructure - no real added value services for users
Summary / Next steps

- C3Grid Prototype running; first project phase ends Sept. 2008

- Next phase: transfer to production environment

- Establishment of a multi-petabyte distributed „climate repository center“ (C3RC) for data storage

- „C3Grid Core Group“ has responsibility to keep this running

- contribution to IPCC AR5 data management

Application for funds in context of next D-Grid Call
Thank you!

further info:
www.c3grid.de
kindermann@dkrz.de
Data provenance
Subject
Issuer
Public Key
Validity
Signature

IdP entityID
SAML
Subject
<Assertion>
...
</Assertion>
>
Shibboleth & VO

Steps:

1. IdP at Institutes ✓
2. Initial SP´s
3. GridShib tests

IVOM/C3-Grid 2007