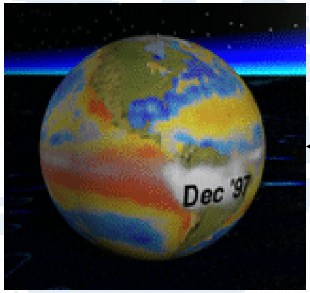
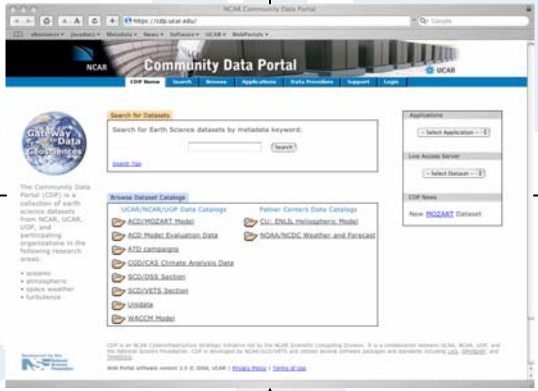
The NCAR Community Data Portal

http://cdp.ucar.edu/

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.





The Community Data Portal (CDP) is a collection of earth science datasets from NCAR, UCAR, UOP, and participating organizations in the following research areas:

- oceanic
- atmospheric
- space weather
- turbulence

Search for Datasets

Search for Earth Science datasets by metadata keyword:

Search

Search Tips

Browse Dataset Catalogs

UCAR/NCAR/UOP Data Catalogs	Partner Centers Data Catalogs
ACD/MOZART Model	CU: ENLIL Heliospheric Model
ACD Model Evaluation Data	NOAA/NCDC Weather and Forecast
ATD campaigns	
CGD/CAS Climate Analysis Data	
SCD/DSS Section	
SCD/VETS Section	
Unidata	
WACCM Model	

Applications

-- Select Application --

Live Access Server

-- Select Dataset --

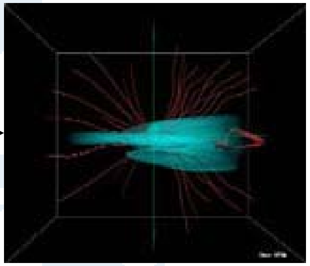
CDP News

New MOZART Dataset

Sponsored by the National Science Foundation

CDP is an NCAR Cyberinfrastructure Strategic Initiative led by the NCAR Scientific Computing Division. It is a collaboration between UCAR, NCAR, UOP, and the National Science Foundation. CDP is developed by NCAR/SCD/VETS and utilizes several software packages and standards including LAS, OPeNDAP, and THREDDS.

Web Portal software version 1.0 © 2006, UCAR | Privacy Policy | Terms of Use



QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

CDP Staff

(VETS: Visualization and Enabling Technologies Section)

Principal Investigator: Don Middleton

Software Engineers: Dave Brown, Mike Burek, Luca Cinquini

Web Designer: Markus Stobbs

Student Assistant: James Humphrey

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Sponsored by the National Science Foundation

Outline

- Introduction
- Architecture
- Describe & demo current functionality:
  - Data catalog browsing
  - Data download
  - Data search & discovery
  - Data aggregation
- Future plans

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

CDP Goals

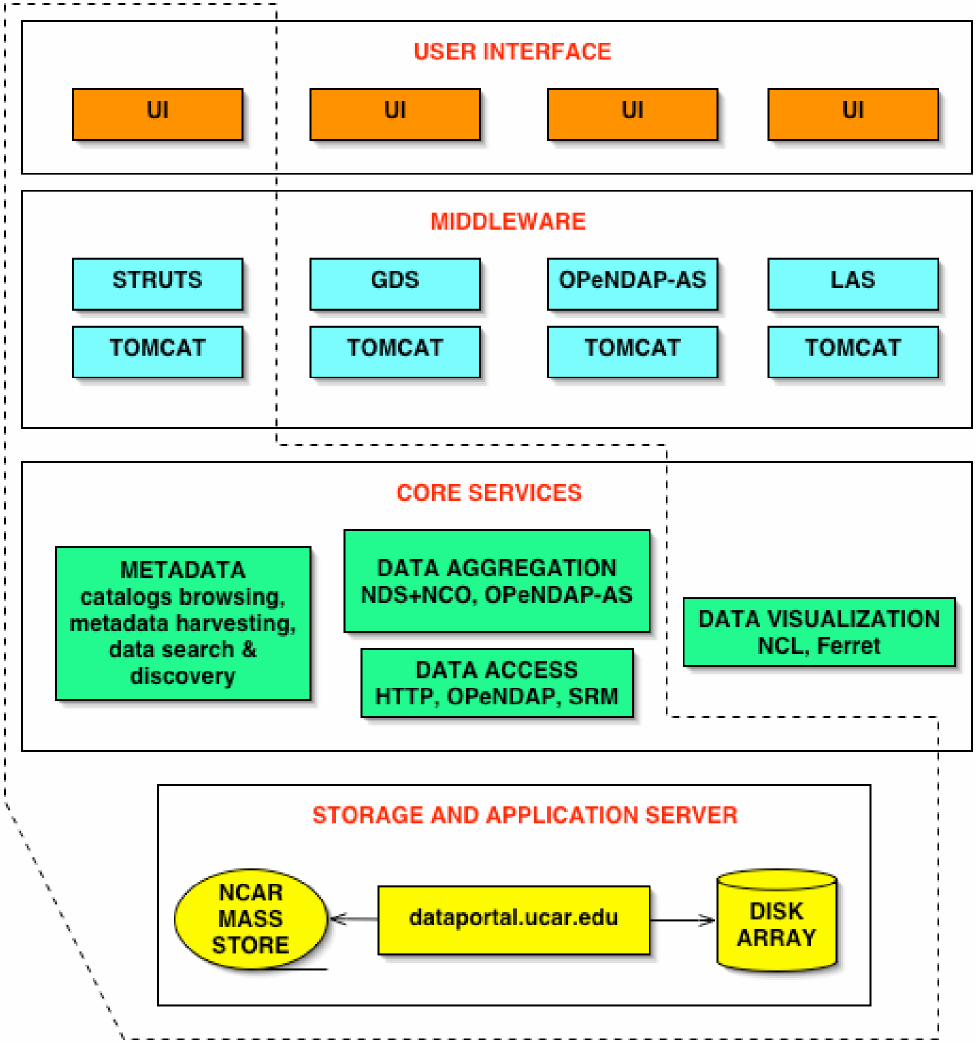
- Develop unified gateway to the large, diverse UCAR/NCAR/UOP data holdings, providing a wide range of data services on these holdings: publishing, browsing, search and discovery, download, remote access, analysis, visualization
- Build the cyberinfrastructure for the integration and support of a broad range of geo-informatic projects within UCAR, thus reducing startup cost and development time
  - Provide physical resources (disk space, computational power)
  - Install, support and integrate non-trivial third-party software packages (Globus/grid environment, OPeNDAP, GRADS, LAS, arcIMS server, etc.) for use by many projects
  - Research and development of reusable components (metadata schemas, digital registration software, aggregation and subsetting of datasets, activity metrics, etc.)

CDP Strategy

- Build unified interface to a distributed, heterogeneous data environment where data is stored at separate locations and managed by different entities
- Collaborate with other UCAR/NCAR/UOP data providers to allow interoperability and promote institution-wide standards; do not take over other groups responsibilities
- Allow for graduated levels of service where data providers choose the extent to which they leverage CDP resources
- Integrate wide range of state of the art technologies from IT realm or geosciences-specific

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

CDP Architecture


USER INTERFACE
UI
UI
UI
UI
MIDDLEWARE
STRUTS
TOMCAT
GDS
TOMCAT
OPeNDAP-AS
TOMCAT
LAS
TOMCAT
CORE SERVICES
METADATA catalogs browsing, metadata harvesting, data search & discovery
DATA AGGREGATION NDS+NCO, OPeNDAP-AS
DATA ACCESS HTTP, OPeNDAP, SRM
DATA VISUALIZATION NCL, Ferret
STORAGE AND APPLICATION SERVER
NCAR MASS STORE
dataportal.ucar.edu
DISK ARRAY

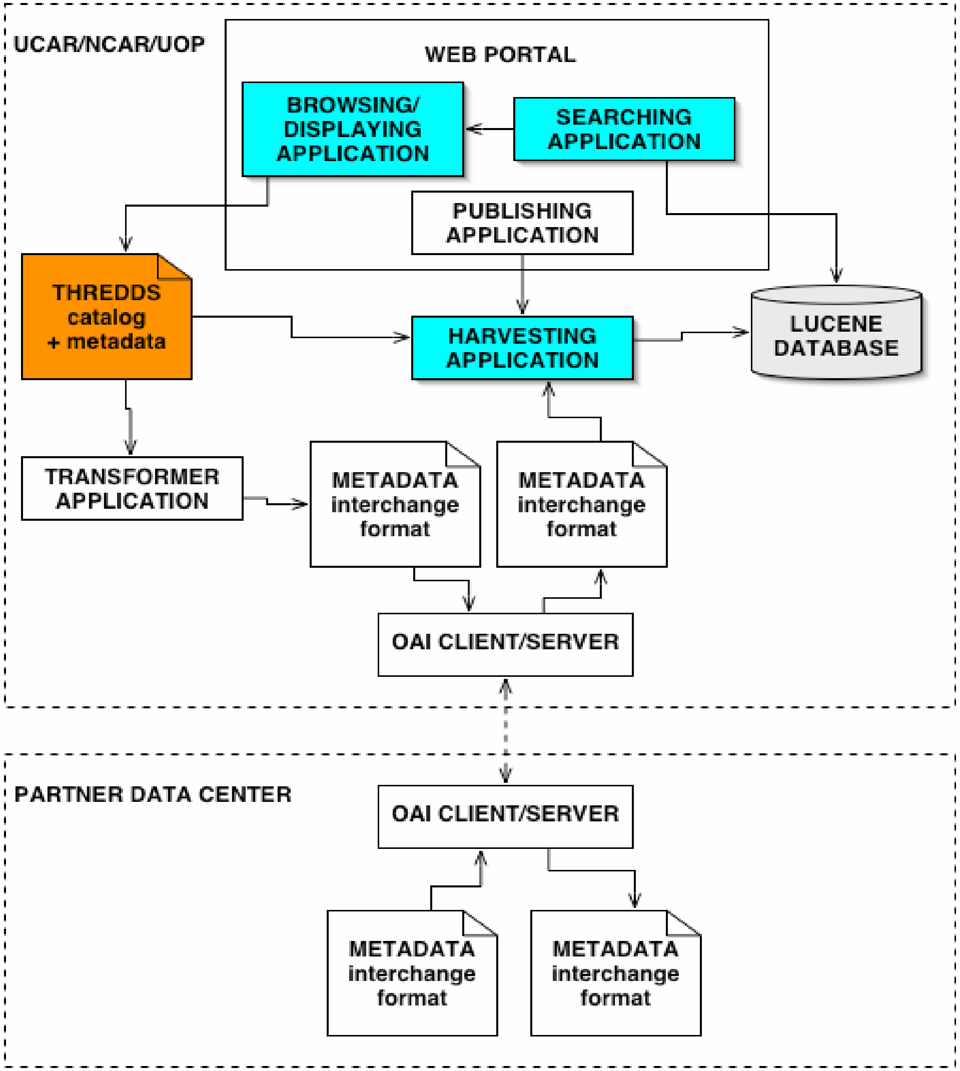
Metadata

- CDP metadata model is based on THREDDS schema:
  - Hierarchical organization of datasets → catalogs browsing
  - Embed/reference descriptive metadata → data search & discovery
- Developed new CDP software components for parsing, harvesting and displaying
- Worked closely with UCAR ITC Data Management Working Group to evaluate/select metadata standards
- Collaborated with Unidata to draft enriched THREDDS metadata (schema version 1.0)
- Data catalogs are XML files served by a web server > distributed, i.e. may be referenced from CDP by URL
- THREDDS v1.0 metadata is mappable to DC, DIF, WMO core (and consequenlt core ISO 19115)

THREDDS catalog example

- <catalog name=“Rainfall Model data catalog”>
    - <service base="http://server.edu/data/" serviceType="HTTPServer" name=”download" />
    - <dataset name=“Rainfall Model” ID=“rain.model” harvest=“true”>
        - <metadata xlink:href=”rain.metadata.xml" metadataType=THREDDS" />
        - <dataset name=“Run 1” ID=“rain.model.run1”>
            - <dataset name=“January 04” ID=“rain.model.run1.200401”>
                - <access serviceName=“download” urlPath=“200401.nc”/>
            - </dataset>
        - </dataset>
        - <dataset name=“Run 2” ID=“rain.model.run2”>
            - ...
        - </dataset>
    - </dataset>
- </catalog>

Metadata Architecture



Sponsored by the National Science Foundation

Dataset-Level Metadata

- Name or title
- Unique identifier
- Short description
- Longer description
- Subject (GCMD keywords)
- Creator (GCMD keywords)
- Publisher (GCMD keywords)
- Project name (GCMD keywords)
- Contributors
- Variables (CF standard names)
- Time coverage
- Space coverage
- Data format (NetCDF, HDF, ...)
- Data size
- Data type (grid, trajectory, radar)
- Access services (HTTPServer, SRM, OPeNDAP, LAS, ...)
- Rights

Demo

- Catalog browsing
- Data download
  - HTTP
  - MSS
- Data search & discovery

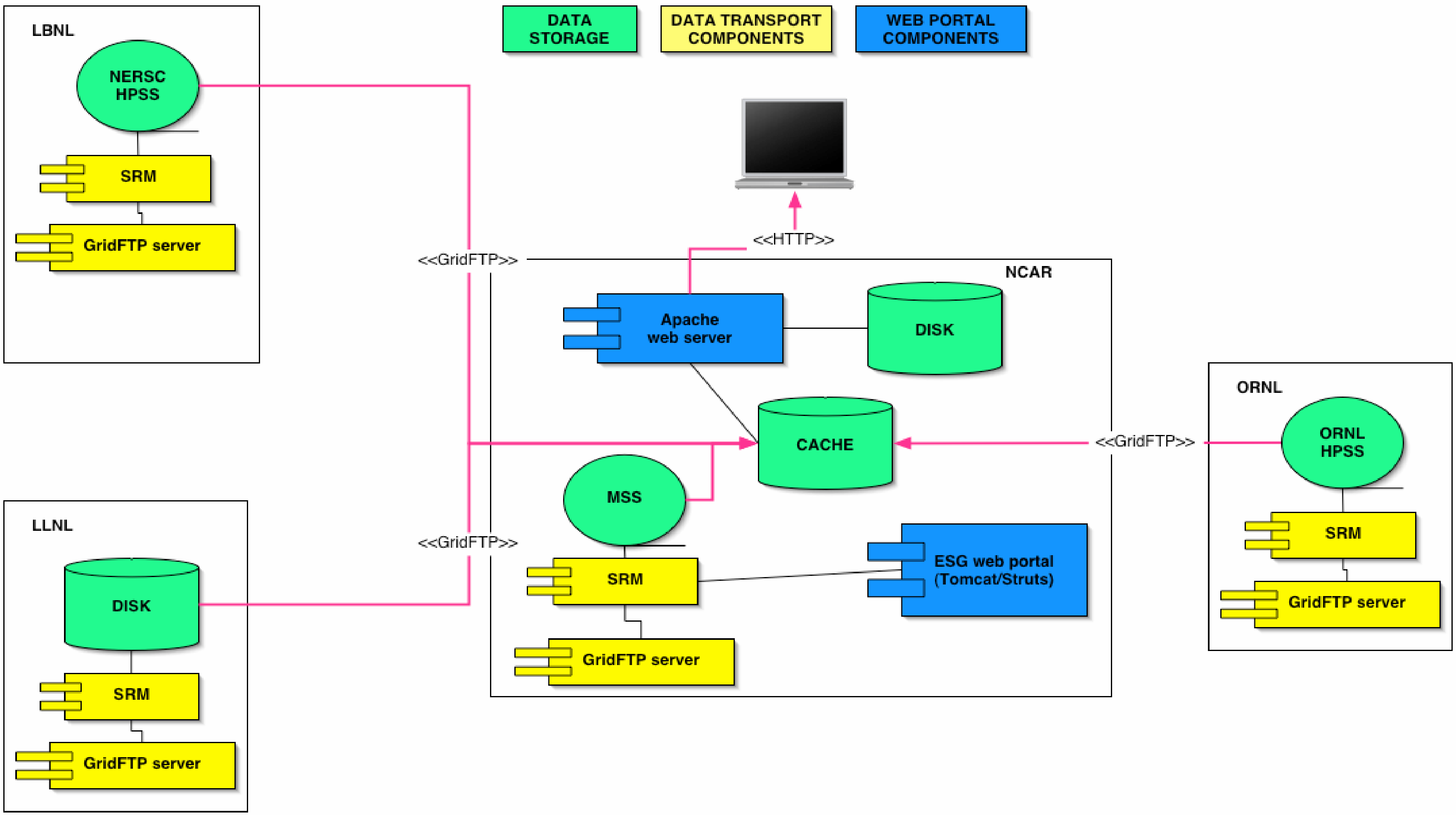
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Data Access

- Online data (on rotating storage):
  - HTTP server: direct download of entire file(s)
  - OPeNDAP: subsetting of single files or aggregated datasets
- MSS data (on tape storage):
  - Use SRM (Storage Resource Manager) developed by ESG/LBNL:
    - Middleware that allows seamless access to data resources whether they are stored on rotating or deep storage
    - File transfer between <u>any</u> deep storage (NCAR MSS, ORNL HPSS, NERSC) and local cache
    - Reliable, high performance transfer between sites via GridFTP
    - Robust, efficient cache management capabilities
  - Requires UCAR Gatekeeper authentication
  - Send email notification when files available on disk cache
- Activity metrics stored in MySQL database

NCAR Community Data Portal UCAR

ESG/CDP data download architecture (deployment diagram)



NCAR/SCD/VETS

Sponsored by the National Science Foundation

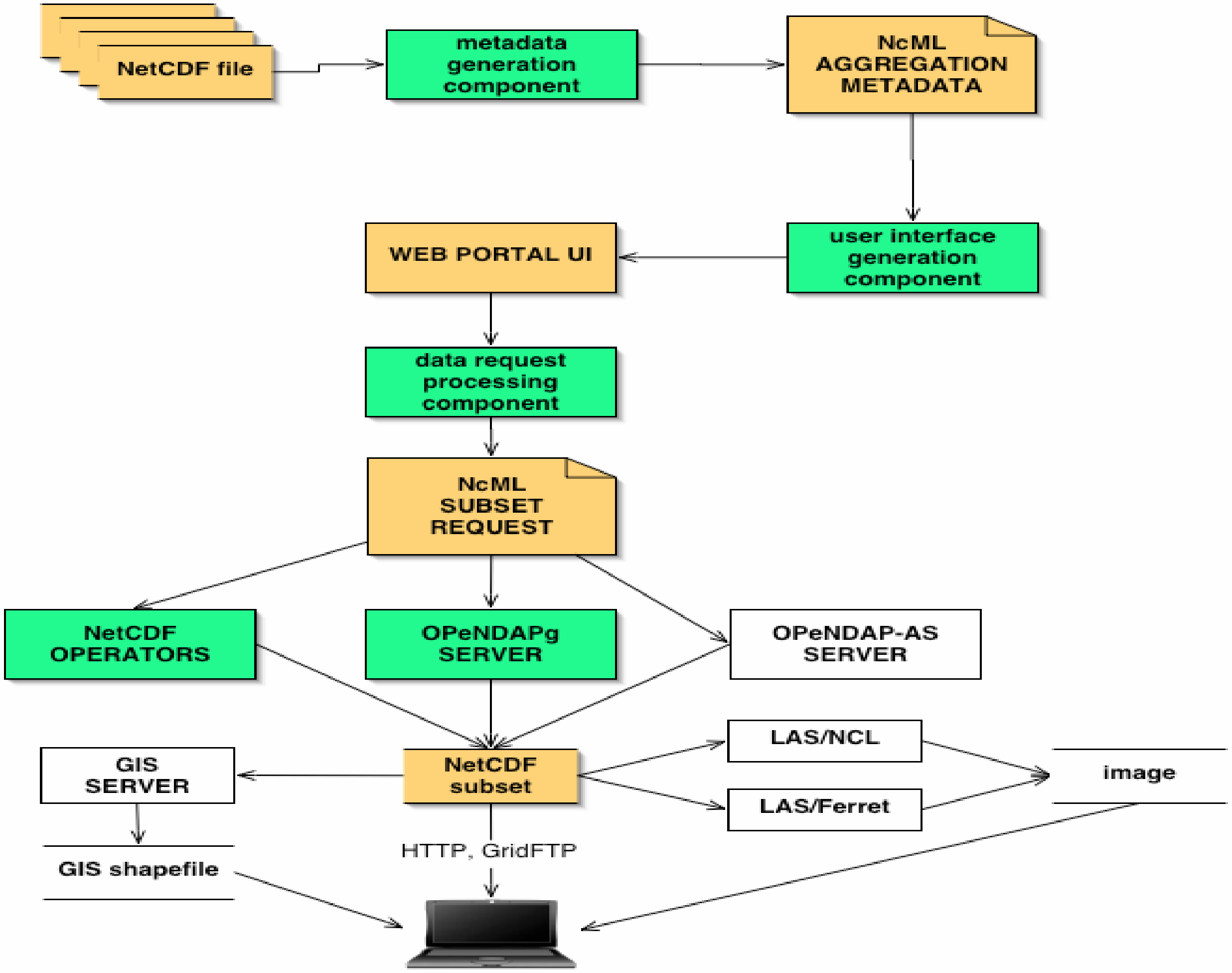
NetCDF Data Aggregation + Subsetting

- Existing technologies: OPeNDAP, OPeNDAP-AS, LAS, NCO
- R&D work that builds upon some of these technologies and provides a modular framework for application-specific integration
- ESG development:
  - Connect OPeNDAP protocol to Grid technologies: high performance data transfer (GridFTP) and GSI (i.e. digital certificates) authentication
    - OpenDAPg, developed by P. Fox & J. Garcia at HAO
  - Publish datasets resulting from multiple levels of aggregation (by variable content and by time coordinate)
    - Develop model for definition of virtual datasets (use NcML!)
    - Develop software for formulating and processing data requests on virtual datasets
    - Modify OpenDAPg to support data aggregation
- CDP requirements:
  - Fast subsetting of aggregated dataset, deliver NetCDF object
  - Simple, intuitive user interface

NetCDF Data Aggregation + Subsetting

- Result: framework for aggregation + subsetting of NetCDF datasets that is modular, flexible and powerful. Different pieces may be combined with existing technologies depending on application requirements
- Workflow:
  1) NcML (NetCDF Markup Language) is used to describe virtual aggregated datasets. Hierarchies of arbitrarily nested NetCDF containers are possible.
  2) Aggregation metadata is used to dynamically generate user interface
  3) User data request is projected from dataset-level to file-level and again encoded in NcML
  4) NcML request document may be processed by pluggable back-end that performs file data extraction and recomposition:
     a) OPeNDAPg (ESG)
     b) NCO (CDP)
  5) Output NetCDF object is delivered to the user (by HTTP, GridFTP, etc.)

NetCDF DATA SERVICES: AGGREGATION AND SUBSETTING



NetCDF file

metadata generation component

NcML AGGREGATION METADATA

user interface generation component

WEB PORTAL UI

data request processing component

NcML SUBSET REQUEST

NetCDF OPERATORS

OPeNDAPg SERVER

OPeNDAP-AS SERVER

GIS SERVER

NetCDF subset

LAS/NCL

LAS/Ferret

image

GIS shapefile

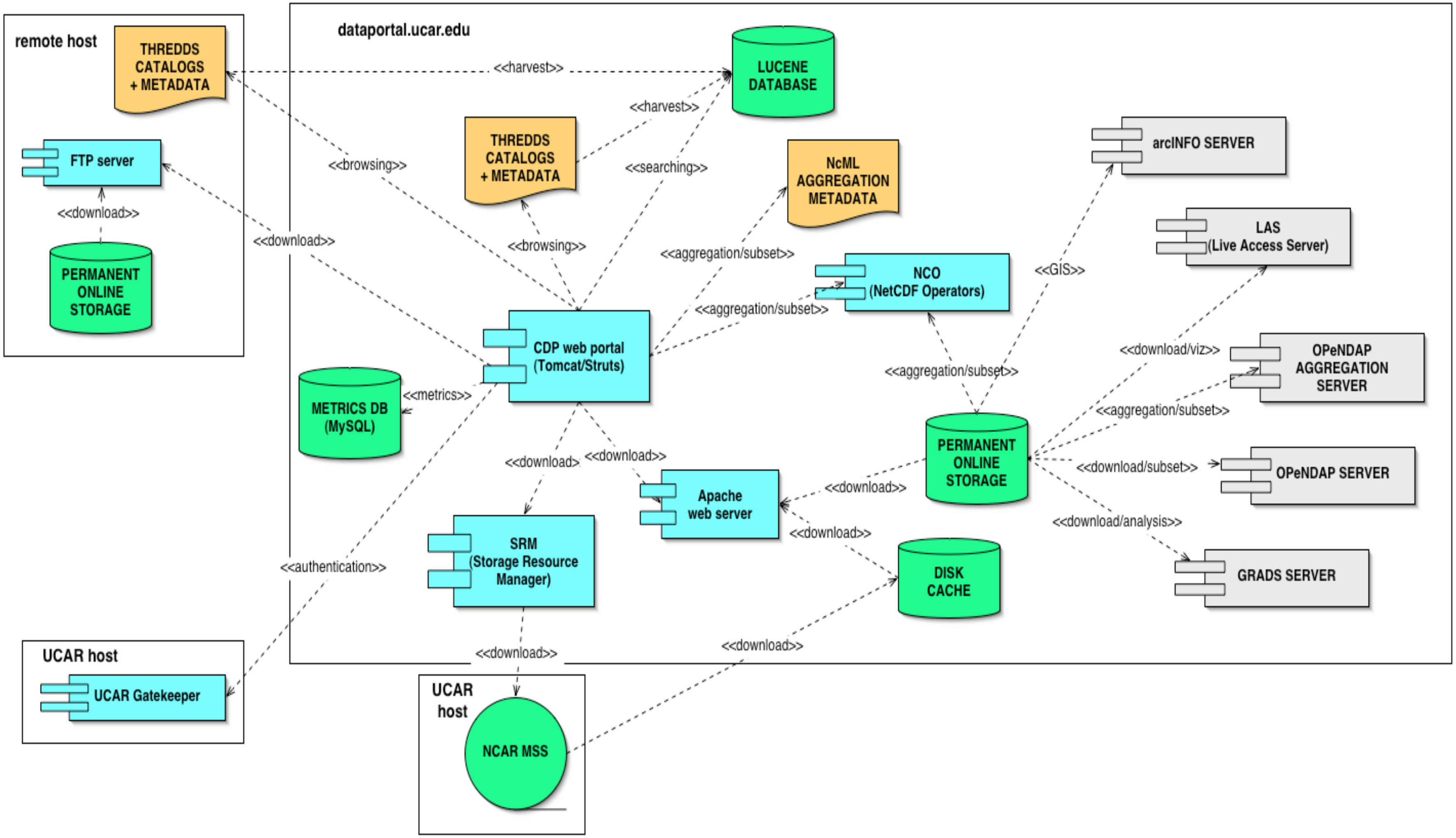
HTTP, GridFTP

Demo

- Data aggregation:
    - WACCM

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

CDP architecture (components diagram)



Sponsored by the National Science Foundation

CDP Top Priorities

- Continue advocacy for institutional participation with DMWG
    - Improve documentation and publishing tools
- Bring portal to production level (stability, monitoring, standard operating procedures)
- Formal user testing and feedback to prioritize future development
- Continue pursuing federation and cooperation with other data centers and projects (NASA GCMD, BADC, WFIS, DLs)
    - Metadata interoperability/conversion
    - Metadata exchange

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

CDP Future Technological Development

- Remote publishing framework
- Increase online storage for high performance data services
- OAI exchange with partner data centers
- Automatic generation of DIF records, publish to GCMD
- Automatic generation of WMO core records, publishing to WFIS centers
- Analyze metrics reports
- Registration and authorization system
- Research and develop visualization services
- Evaluate SRB (Storage Resource Broker) for MSS download

CDP collaborations and acknowledgements

- SCD/DSG: thanks for supporting the hardware!
- SCD/DSS: metadata and data services
- SCD/MSS: online access to MSS
- ESG (including CGD, HAO): shared development, hosting environment, technologies
- Unidata: joint development of NcML, collaborated on THREDDS search and discovery metadata
- DLESE, BADC, GCMD, FWIS: export or exchange (via OAI) of metadata documents for cross-institutional searches
- COLA: provide remote data services through GRADS
- Many data providers across UCAR/NCAR/UOP and others: ACD, ATD, CGD (CAS, PCM, CCSM), JOSS, SCD (DSS, VETS), Unidata, WACCM and CU/ENLIL
- GridBGC: shared development
- GIS: NetCDF to GIS conversion services
- GO-ESSP: sharing information and technologies
- NOMADS: undergoing exploratory collaboration

Appendix: Interoperating with GCMD

- Why not rely completely on GCMD portal to discover data?
  - Because GCMD only provides search and discovery of data, while CDP aims at building a full integrated environment for search, browsing, dowload, analysis and visualization
  - NCAR cannot rely on another institution to provide access to its data
  - GCMD is a central metadata repository ("push" model), while community is evolving towards distributed, cooperating centers
- Why not adopting DIF as metadata standard? It was carefully considered, but:
  - DIF provides dataset-level description, not direct file access
  - DIF, THREDDS play a different role
  - DIF is not an open standard mantained by the community
  - Could embed DIF records within THREDDS catalogs, but would result in duplication and possible inconsistency of metadata
- ... but CDP will interoperate with GCMD and other data centers!