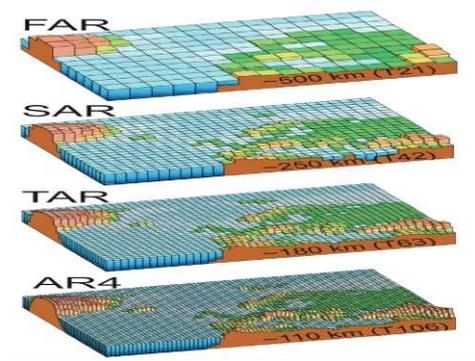


ESG Data Node

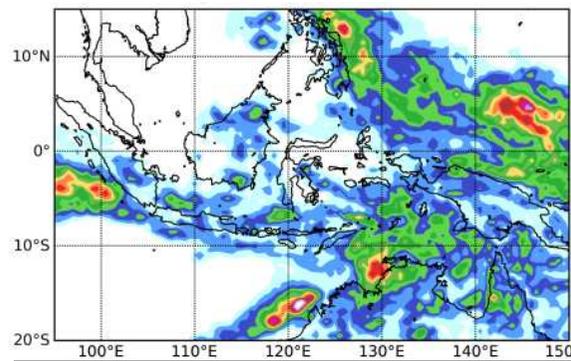


Robert Drach, Gavin Bell, and Dean N. Williams  
On behalf of the **Earth System Grid Center for Enabling Technologies (ESG-CET) Team**  
Presentation at the  
**2009 Global Organization for Earth System Science Portal (GO-ESSP) Workshop Agenda**  
Hamburg, Germany  
October 7, 2009

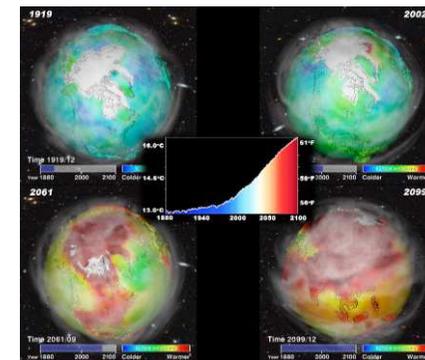
Multi-Model IPCC Assessment Reports



Satellite Precipitation Observations



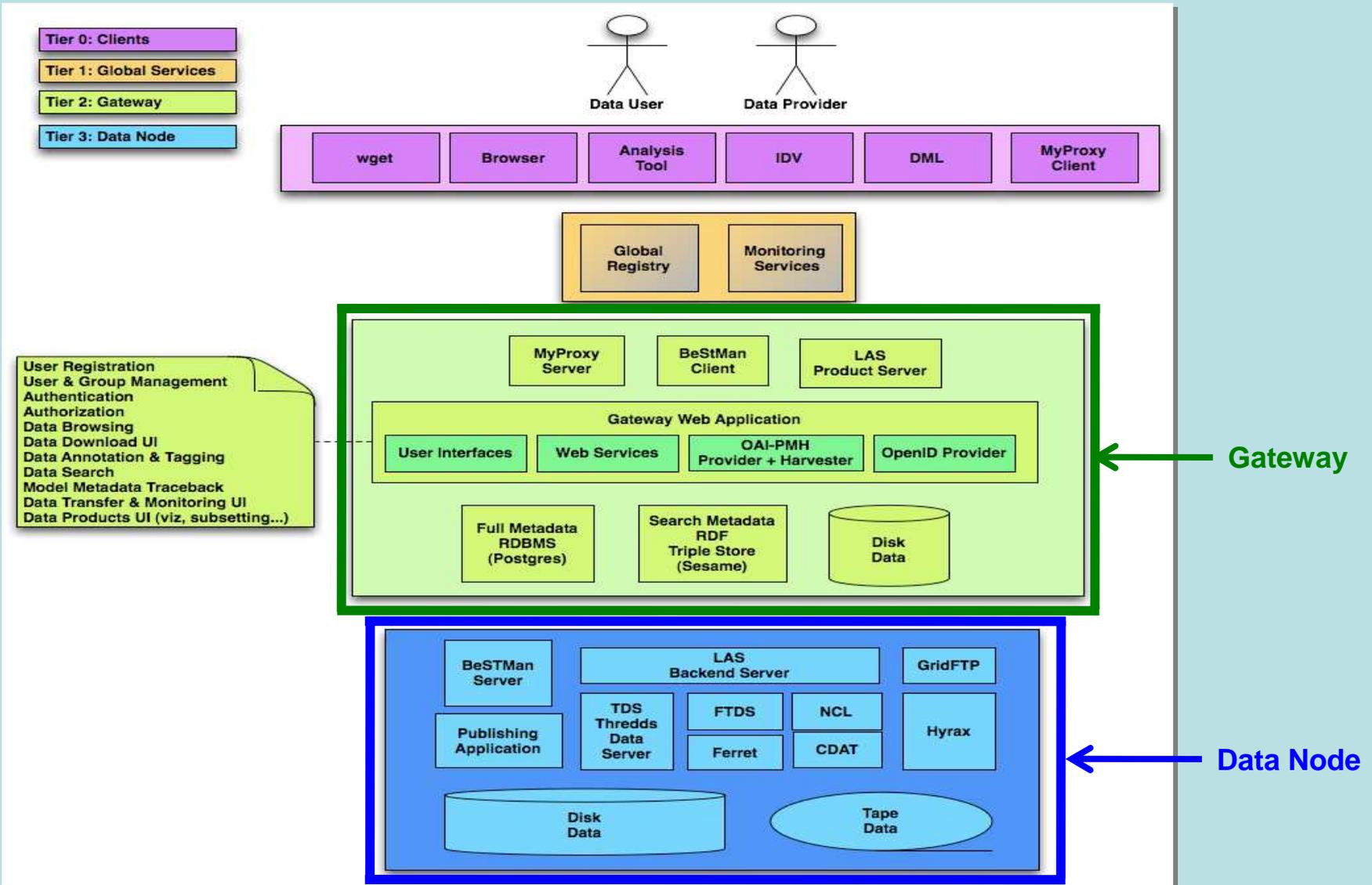
Model Ensembles



# ESG-CET Architecture



Center for Enabling Technologies





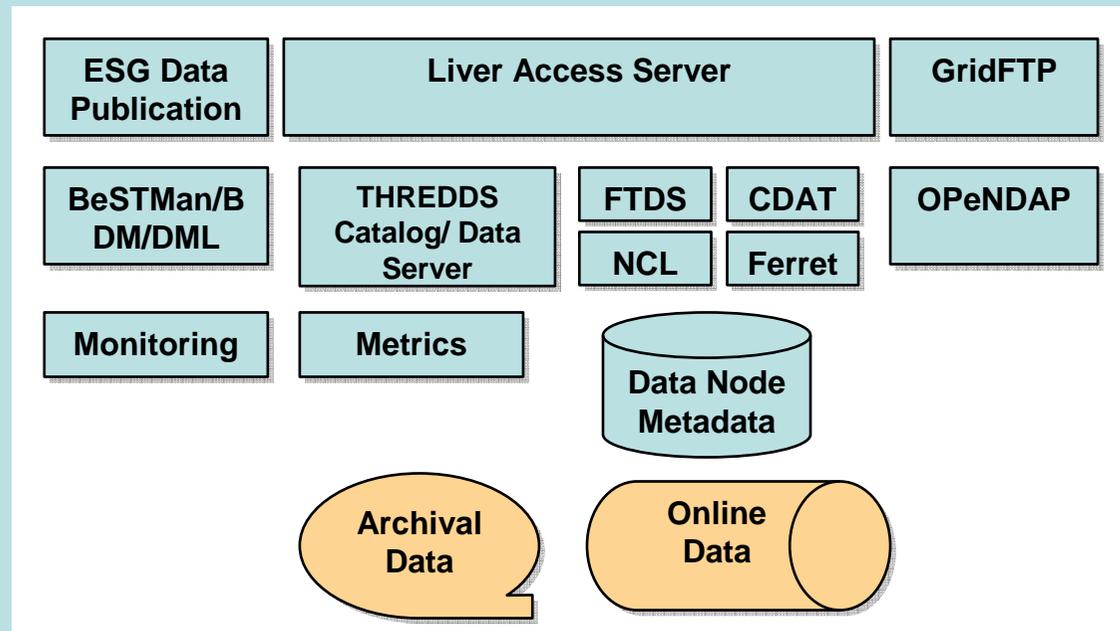
- λ **Federated architecture**  
Federation is a virtual trust relationship among independent management domains that have their own set of services. Users authenticate once to gain access to data across multiple systems and organizations
- λ **Gateways**
  - ➔ Where data is discovered, requested
  - ➔ Portals, search capability, distributed metadata, registration and user management
  - ➔ May be customized to an institution's requirements, topical focus
  - ➔ Fewer sites
  - ➔ Initially **PCMDI**, **NCAR**, ORNL, BADC, eventually GFDL, ANU
- λ **Nodes**
  - ➔ Where data is stored and published
  - ➔ Data may be on disk or tertiary mass store
  - ➔ Each data node can publish to any gateway (facilitates topical gateways)\*
  - ➔ Data reduction/analysis\*
  - ➔ Possible minimalist deployment without services\*
  - ➔ Anticipate ~15 data nodes for CMIP-5, many others have expressed interest
- λ **Sites**
  - ➔ A site can be both a gateway and a data node

## Data Nodes Architecture



Center for Enabling Technologies

- λ **Data publication utilities**
  - ➔ Esgpublish / esgunpublish
  - ➔ Esgpublish\_gui
- λ **Query Data Node / Gateway**
  - ➔ Esglist\_datasets
  - ➔ Esgquery\_gateway
- λ **THREDDS Data Server (TDS): catalog, http download, OPeNDAP services**
- λ **Live Access Server (LAS): Product services**
  - ➔ Front-end to CDAT, Ferret, NCL
- λ **GridFTP: secure high performance download**
- λ **DataMover-Lite (DML): front-end to GridFTP**
- λ **MyProxy client: X509 proxy certificate generation**

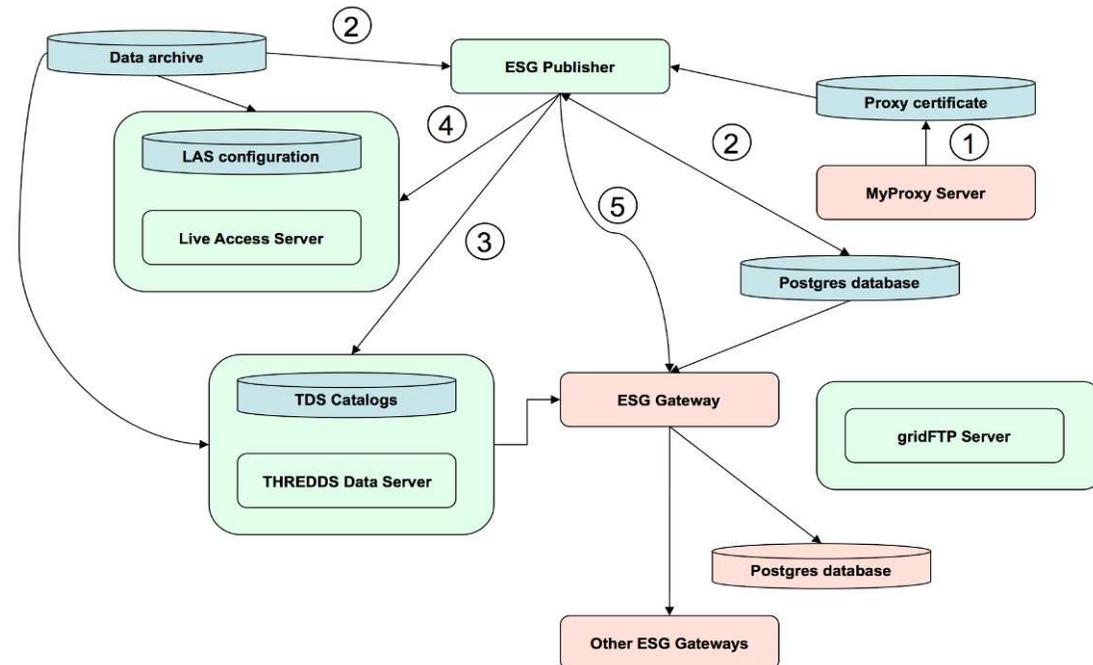


## Publication Flow



Center for Enabling Technologies

1. Data publisher obtains a proxy certificate
  - ➔ Associates files to datasets
  - ➔ Caches file metadata in Data Node database
2. Data Publisher scans a set of files
  - ➔ Associates files to datasets
  - ➔ Caches file metadata in Data Node database
3. Generate THREDDS re-initialize the TDS
4. Re-initialize LAS based on TDS Catalog
5. Publish to Gateway
  - ➔ Contact Gateway via web service
  - ➔ Gateway harvests new TDS Catalogs



**Green** – Data Node components  
**Blue** – Data and configuration files  
**Red** – Gateway services that interact with the publisher



- λ Considerable recent effort to package installation
- λ Support Redhat Linux EL5, CentOS
  - ➔ Installation script
- λ Virtual Machine (VM)
  - ➔ Installation script
  - ➔ Complete VM with required libraries pre-installed
  - ➔ Xen (planned)
- λ Resources
  - ➔ <http://esg-pcmdi.llnl.gov/internal/software-packaging-deployment-and-maintenance/>
  - ➔ Esg-node-dev@lists.llnl.gov



- λ **INI format text file, typically one per Data Node**
- λ **Global parameters:**
  - ➔ Database connection, TDS configuration, Gateway locations, ...
- λ **Per-project configuration**
  - ➔ Directory structure, dataset identifiers
  - ➔ Search fields defined
  - ➔ Predefined search fields
    - ➔ Project, experiment, model, run\_name, variable
  - ➔ Format strings: dynamic pattern substitution
- ➔ Maps: Enable flexible dataset ID generation

```
time_frequency_map = map(time_frequency_short : time_frequency)
3h | 3hourly
da | daily
fixed | monthly
```



- λ **Initial release milestone in Oct. '09**
  - ➔ Support for CMIP5
  - ➔ Browse context-sensitive search
  - ➔ Direct download, multi-file download with wget or DataMover-Lite
- λ **Prototype Gateways:**
  - ➔ PCMDI and NCAR
- λ **Data Node deployments underway:**
  - ➔ PCMDI, NCAR, ORNL, GFDL, NASA JPL and Goddard, BADC, ANU, University of Tokyo/JAMSTEC, DKRZ
- λ **Versioning: tracking multiple versions of datasets / files**
  - ➔ Main use case is to support notification
  - ➔ Track data errors / corrections
- λ **Metrics**



- λ Dataset replication
- λ Server-side analysis
- λ Support for broader range of data sets
  - ➔ Plugin architecture for customized project handlers
- λ Gateways:
  - ➔ ORNL, GFDL, BADC, ...
- λ Data Node:
  - ➔ Other CMIP5 modeling centers