LLNL-PRES-471684

ESGF
Earth System Grid Federation

Earth System Science Portals: Earth System Grid Federation (ESGF)

Dean N. Williams on behalf of the ESGF Community

Federation and Integrations of Data from Multiple Sources

SCExSS-OC◆ October 11, 2011

Office of Science
U.S. DEPARTMENT OF ENERGY

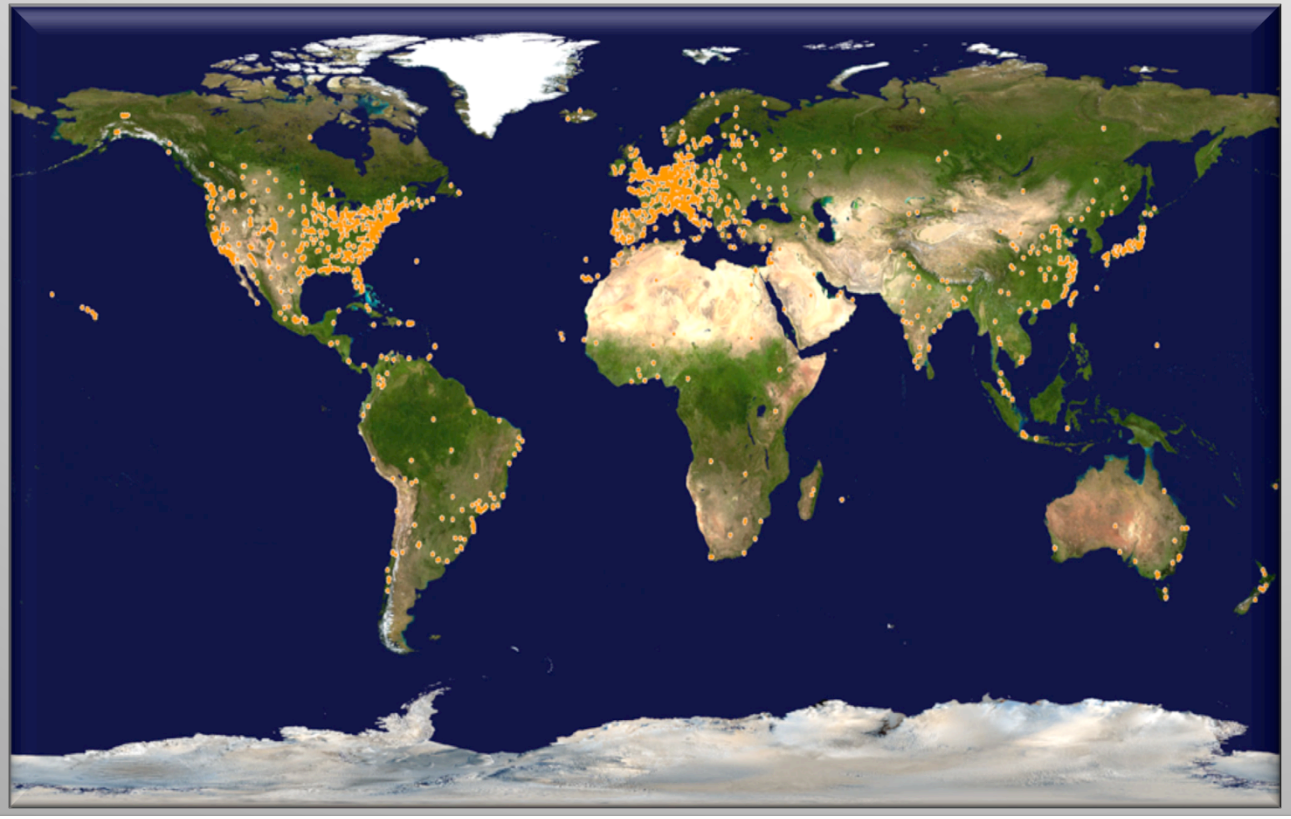
SciDAC
Scientific Discovery through Advanced Computing

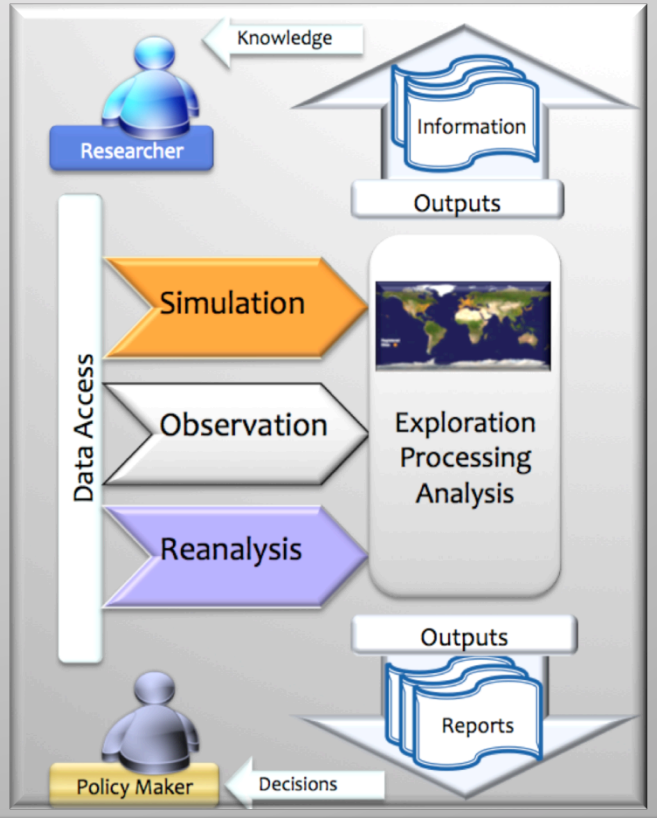
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

Earth System Grid Federation
ESGF

Earth System Grid Federation (ESGF):
Coupled Model Intercomparison Project

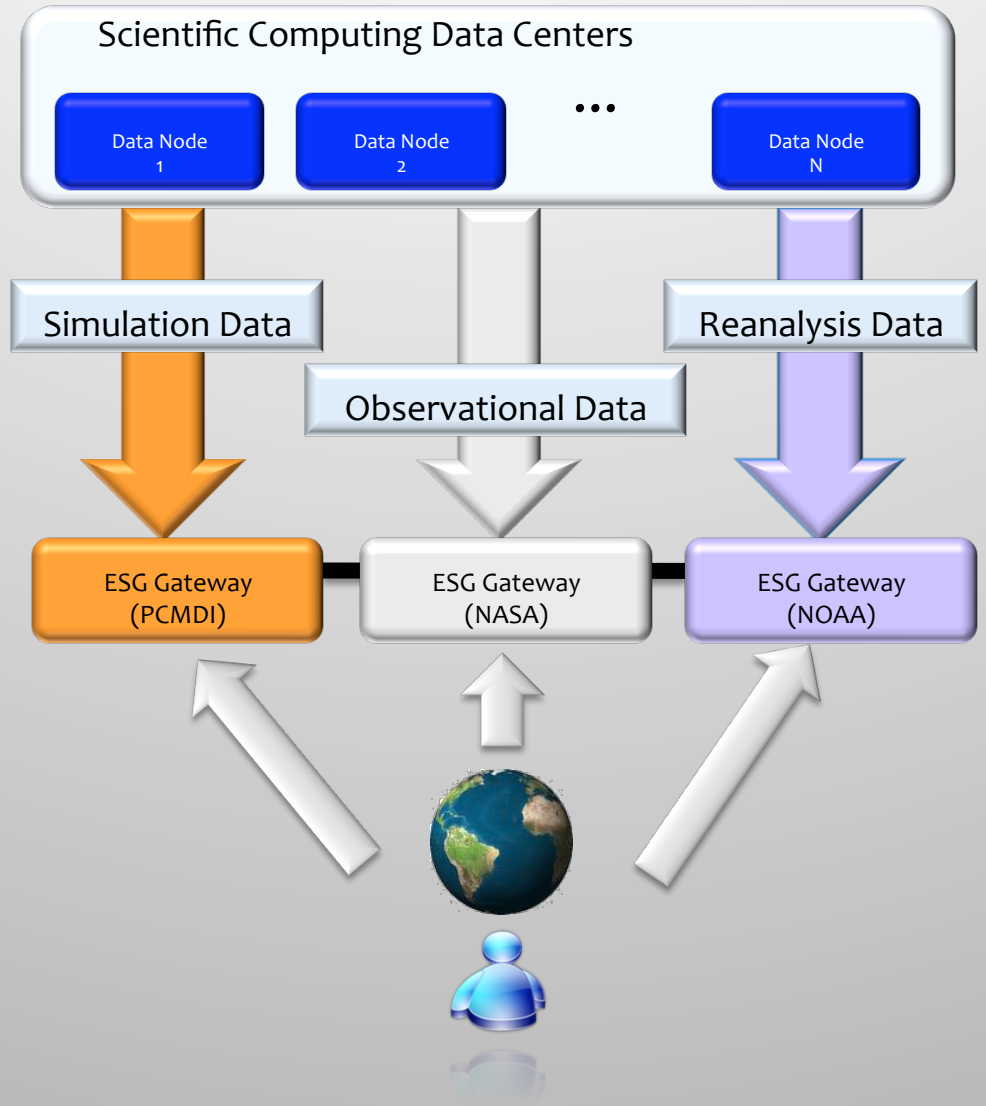
- ESGF is a free, open consortium of institutions, laboratories and centers around the world that are dedicated to supporting research of Climate Change, and its environmental and societal impact
- Historically originated from Earth System Grid (ESG) project, expanded beyond its constituency and mission to include many other partners in the U.S., Europe, Asia, and Australia
- Groups working at many projects: ESG, Earth System Curator, Metafor, Global Interoperability Program, Infrastructure for the European Network for Earth System Modeling, and many more
- U.S. funding from DOE, NASA, NOAA, NSF

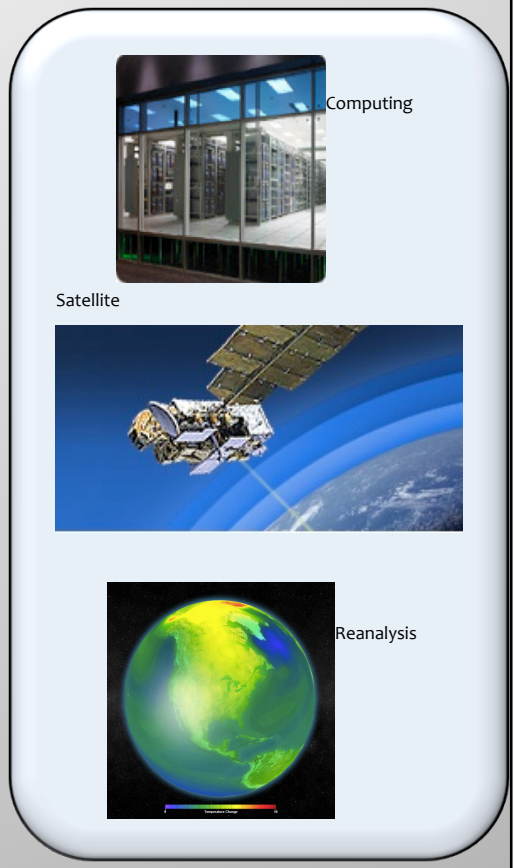




ESGF development of climate data management, access and analysis

- Too much data to move, must leave in place
- Promote sharing of knowledge, software and tools among partners
- Define APIs and protocols for interoperability among data centers
- Collaborative development of some software components
- Deployment of common software infrastructure

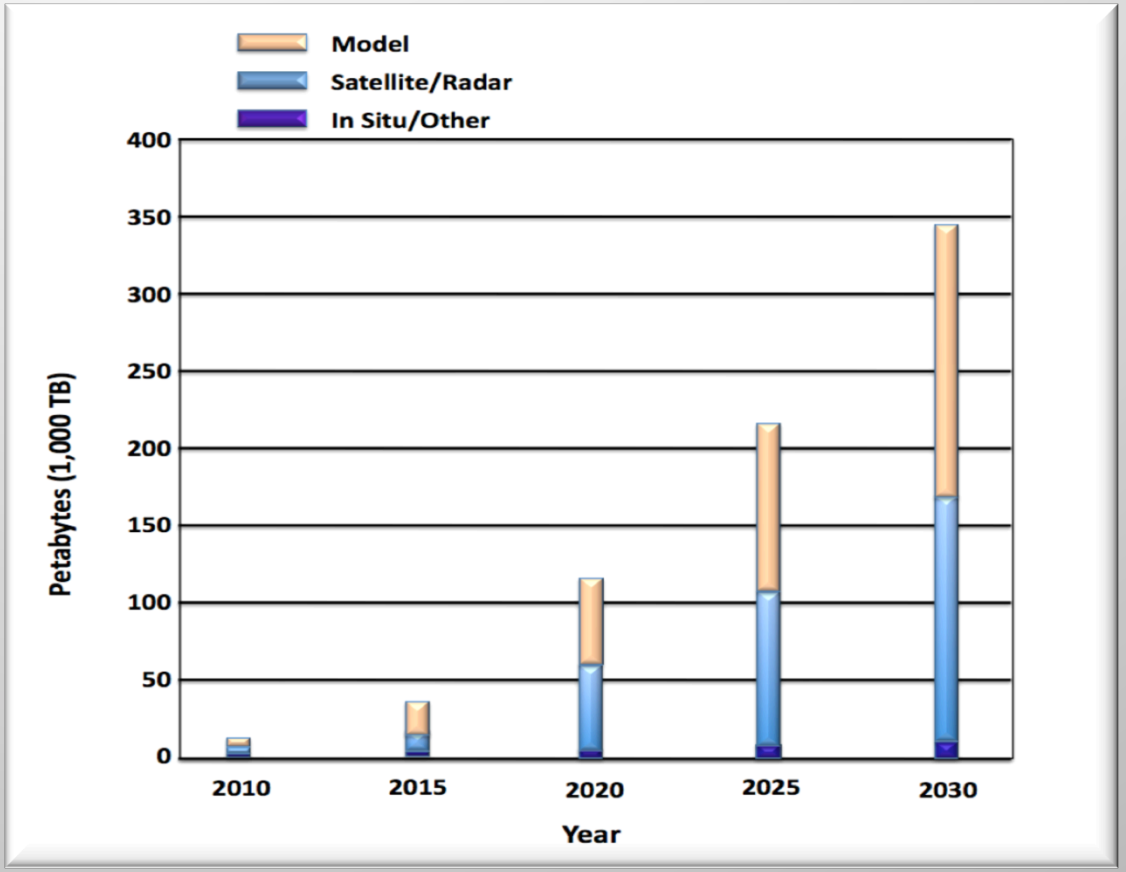




Community climate data storage requirements expands rapidly every five years as computer capabilities increase and climate simulations become more complex

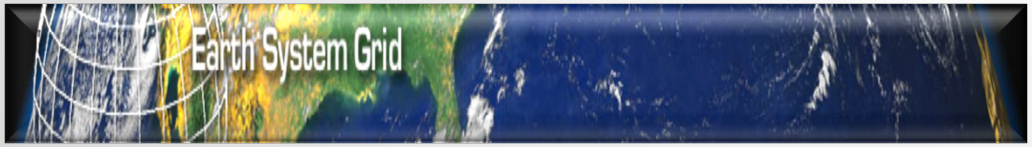
J.T. Overpeck, G.A. Meehl, S. Bony, D.R. Easterling
"Climate Data Challenges in the 21st Century", Science, 11 February 2011, Vol 331, 700 (2011); DOI: 10.1126/science.1197869"

- Rapidly expanding worldwide global climate data holdings
- Local and remote servers with web-based and analysis tool access.
- >1 Tbps networking must fit in with the timelines shown in the graph
  - End-to-end monitoring for centers and individual desktop transfers and remote access and server-side analysis
- Climate network mailing list: climatedata@es.net



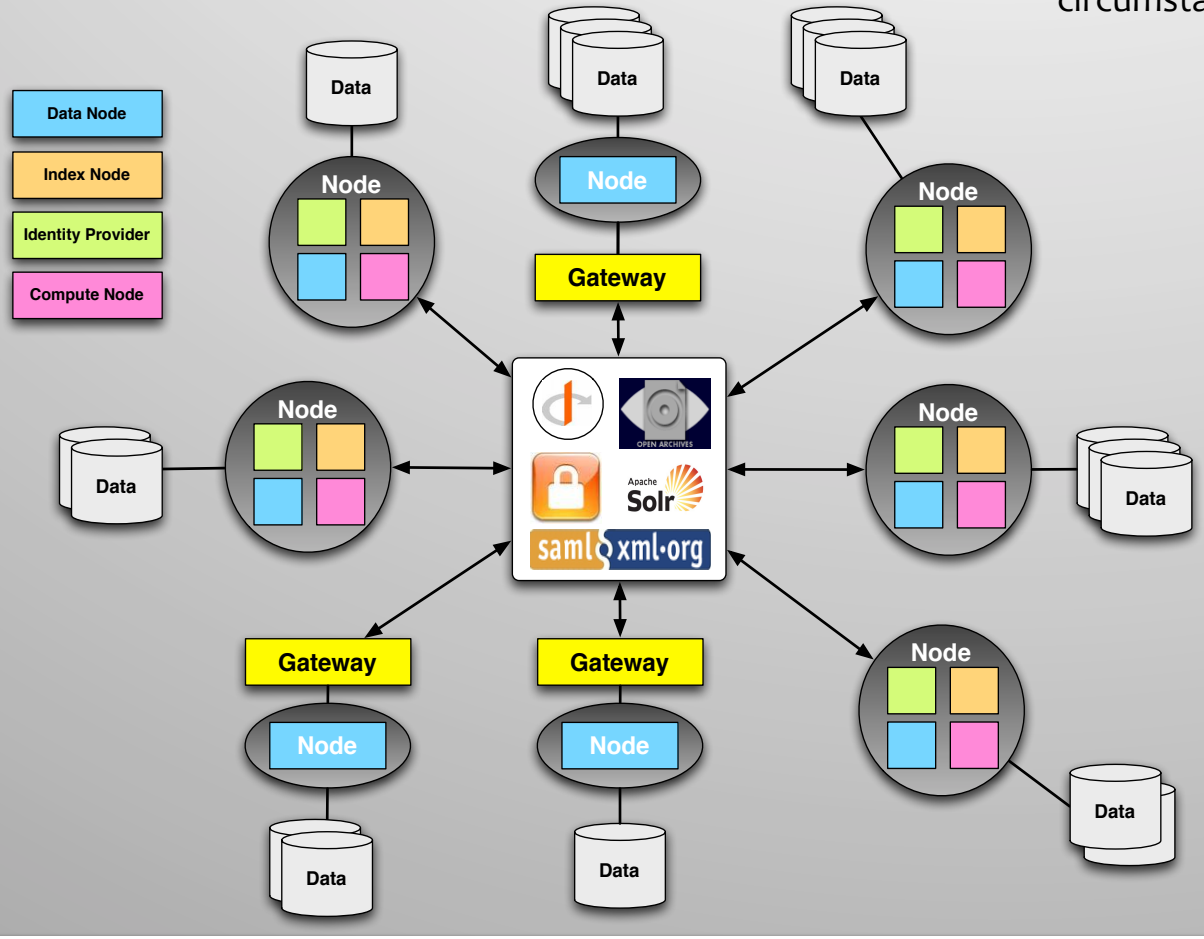
Transfer Rate	Time to Transport 1 TB of Data	Time to Transport 33.3 TB of Data	Time to Transport 1 PB of Data	Time to Transport 3 PB of Data
10-Gbps	13.65 minutes	7.6 hours	9.7 days	29.1 days
100-Gbps	81.9 seconds	45.5 minutes	23.3 hours	2.9 days
1-Tbps	8.19 seconds	4.55 minutes	2.33 hours	6.7 hours

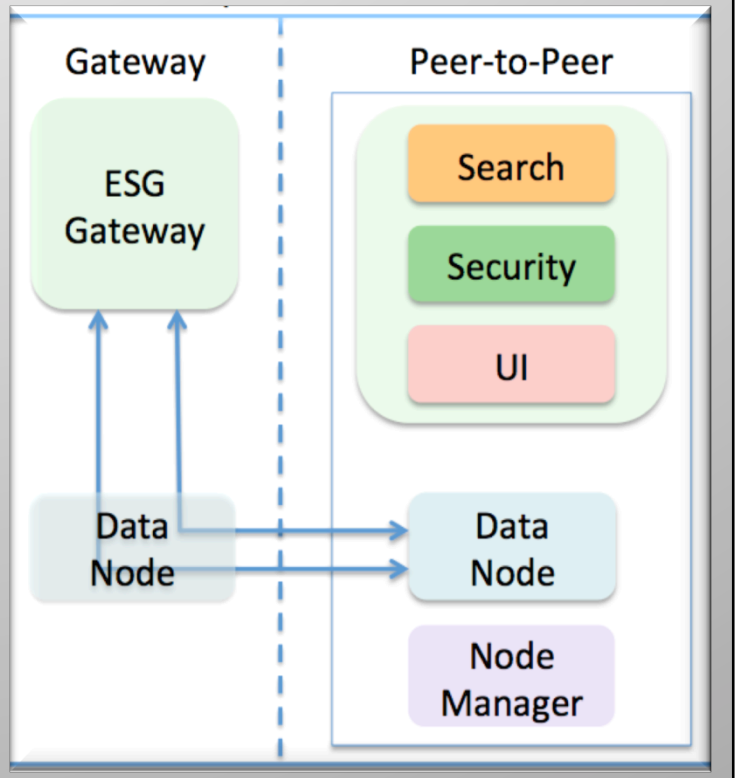
Balanced Ecosystem: Overall Architecture Design



Gateway-data node (client-server): A more traditional model, based on a specialized gateway application that acts as a broker towards services provided by a data node

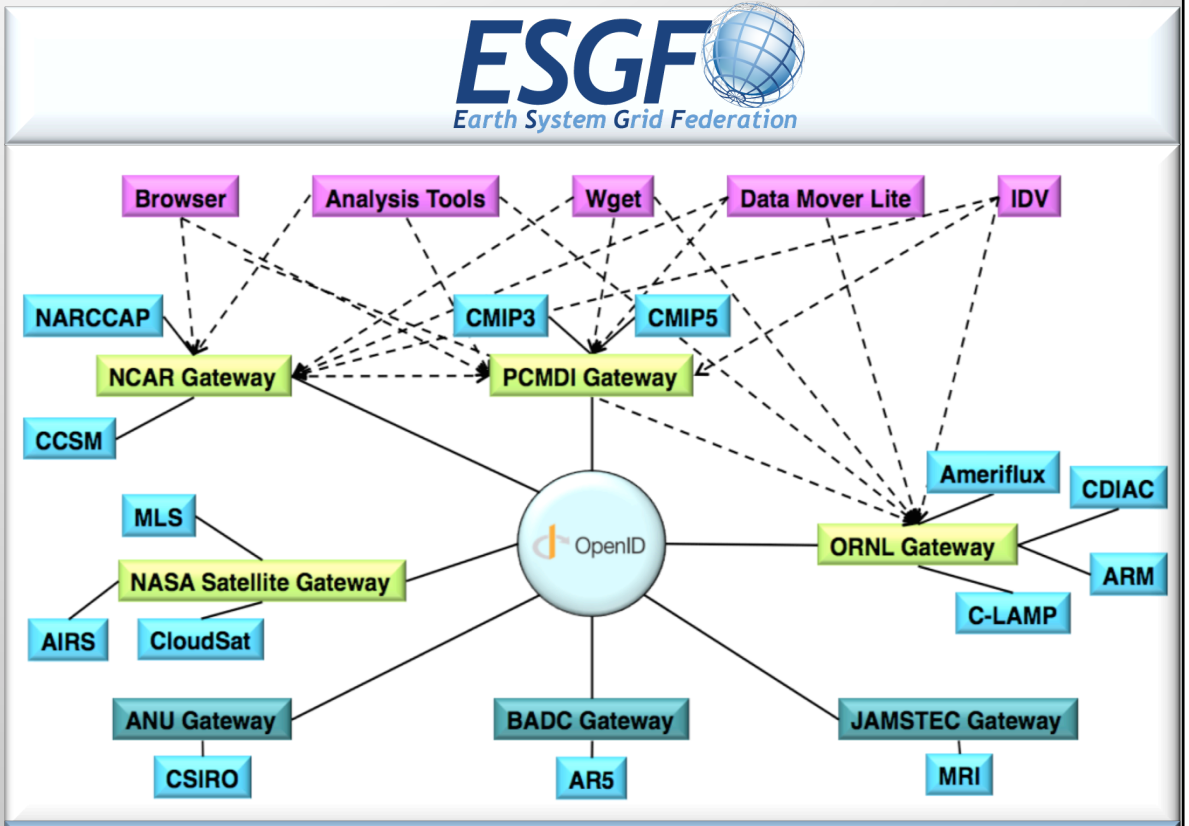
Peer-to-Peer (P2P): An innovative paradigm in which all participating sites interact as equal partners, can be flexibly configured to expose different sets of services, and can act as consumers or providers of services depending on circumstances





The ESGF is a distributed data archival and retrieval system

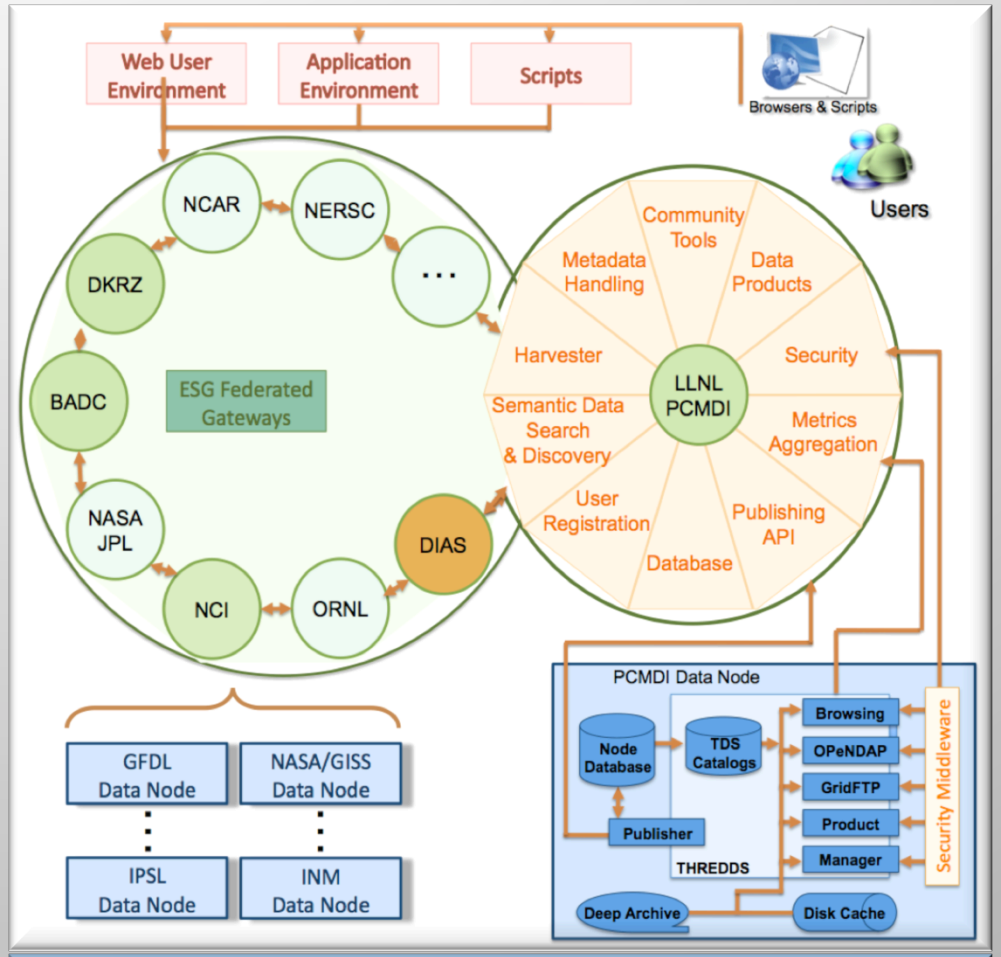
- Distributed and federated architecture
- Support discipline specific Gateways
- Support browser-based and direct client access
- Single Sign-on
- Automated GUI-based publication tools
- Full support for data aggregations
  - A collection of files, usually ordered by simulation time, that can be treated as a single file for purposes of data access, computation, and visualization
- User notification service
  - Users can choose to be notified when a data set has been modified



Large amounts of data movement and management are needed for the climate community.

Networking the climate science community for large-scale data and science discovery

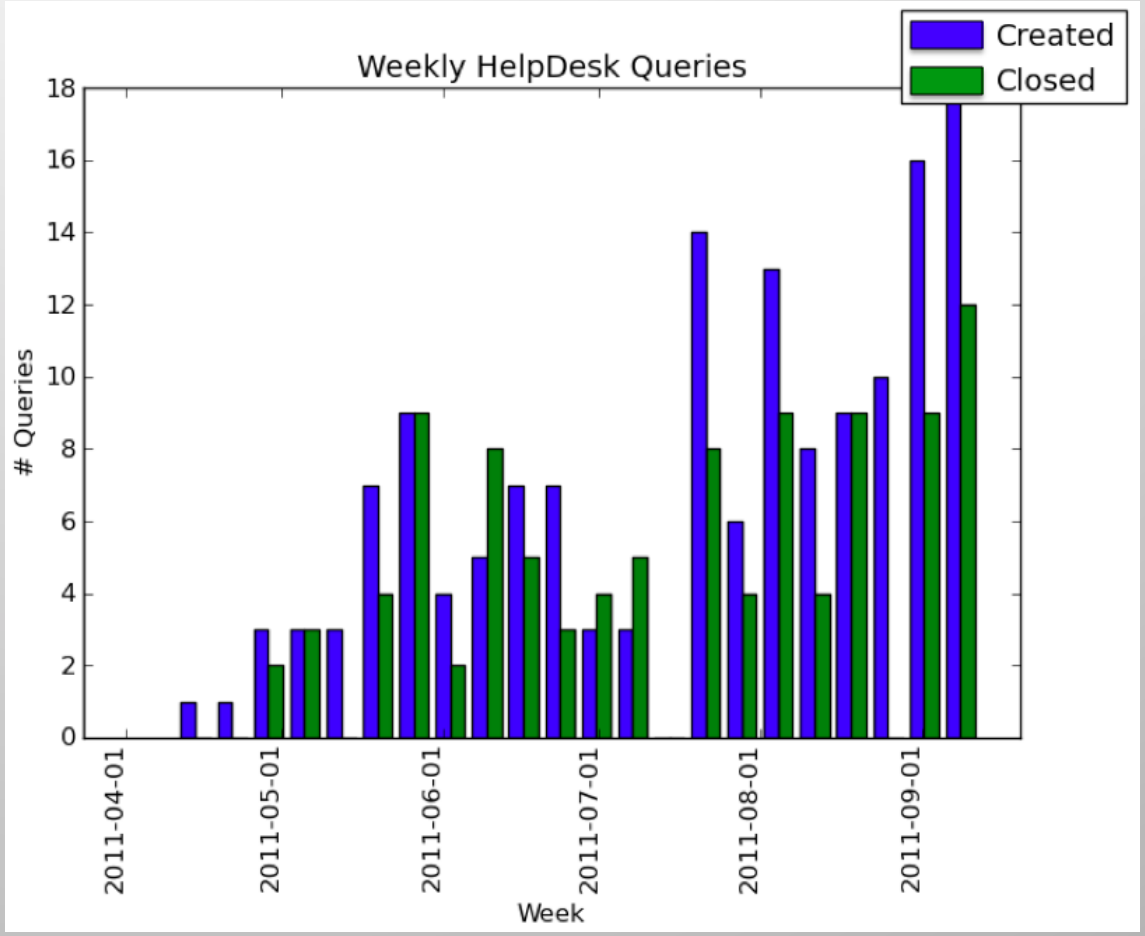
- CMIP = Coupled Model Intercomparison Project
  - Phase 1: Idealized simulations of present-day climate
  - Phase 2: Idealized simulations of future climate changes
  - Phase 3: More realistic simulations
- CMIP 5 multi-model archive expected to include
  - 3 suites of experiments
  - 25 modeling centers in 19 countries
  - 58 models
  - Total data, ~10 PB
  - Replica 2 – 3 PB
- Global distribution
- Timeline fixed by IPCC (2011 – 2013)
- Wide adoption by the Climate community: CMIP, CCSM, CSSEF, ARMBE, Obs4MIPs, TAMIP, NCPP, NARCCAP, NCA, etc.



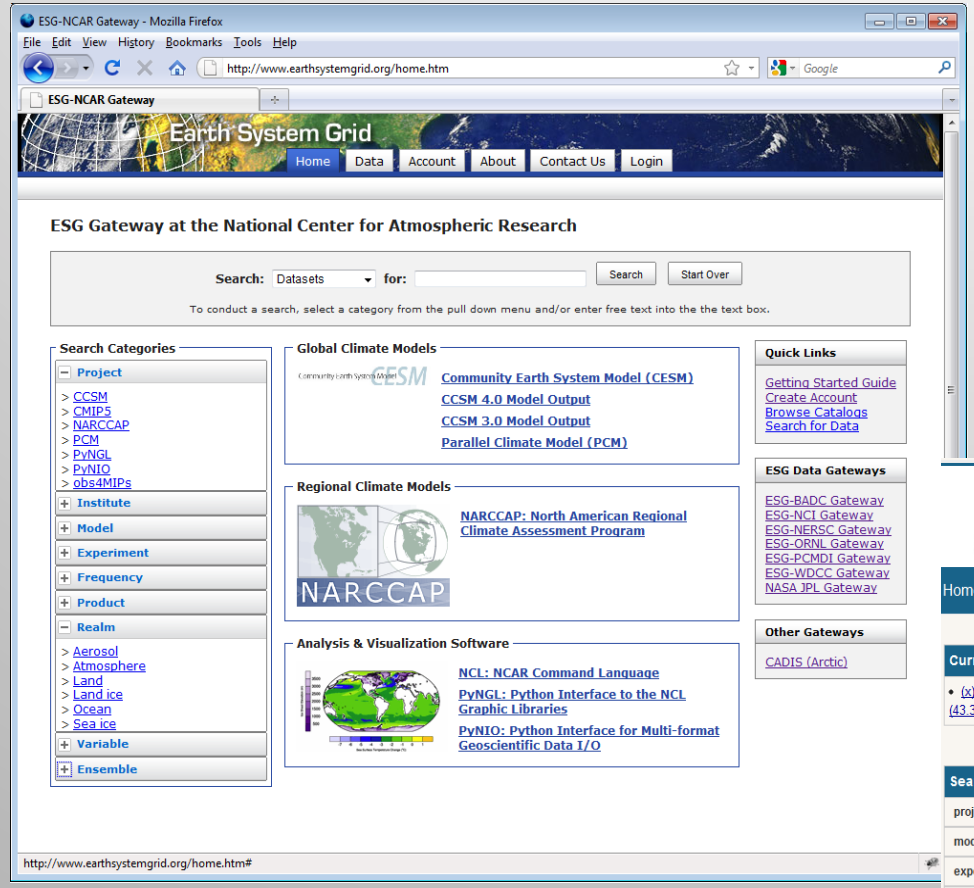
High performance networks are needed for ESGF to replicate data worldwide and to move data for large-scale analysis.

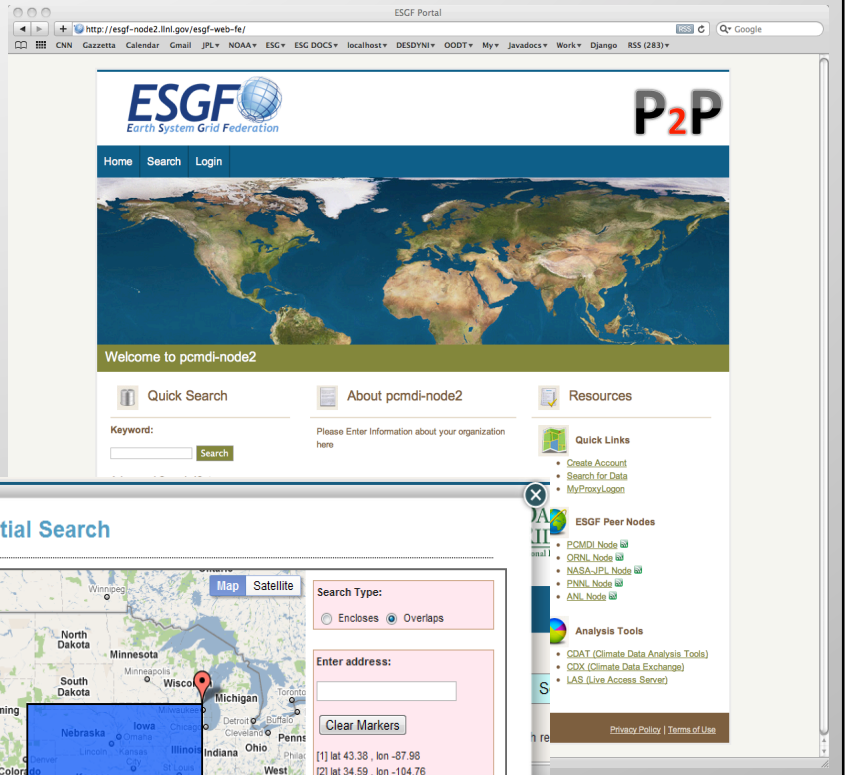
ESGF Help Desk Weekly Traffic

- About half of these queries come directly to cmip5-helpdesk@stfc.ac.uk and the other half to the esg-support mailing list esg-support@earthsystemgrid.org
- Questions that are resolved are place on the ESGF FAQ list
- Scientists at BADC, PCMDI, and DKRZ are charged with addressing CMIP5 data questions
- Technical staff at BADC, PCMDI, DKRZ, NCAR, and JPL are charged with addressing ESGF system questions

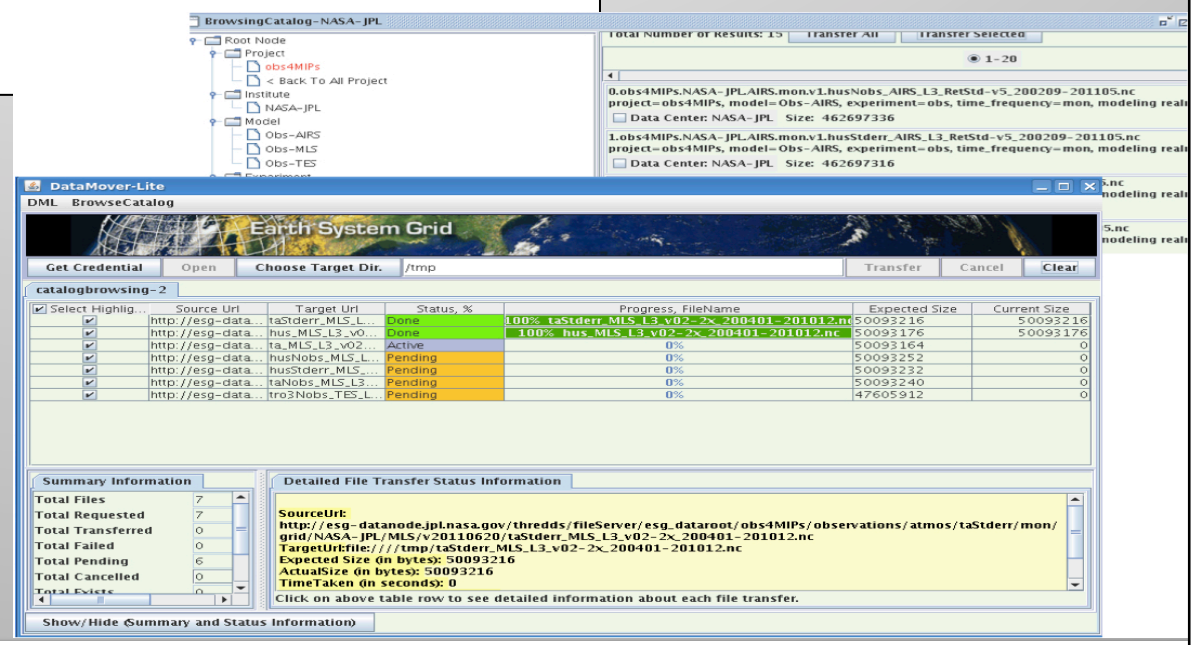


Gateway and Peer-to-Peer Web Front End



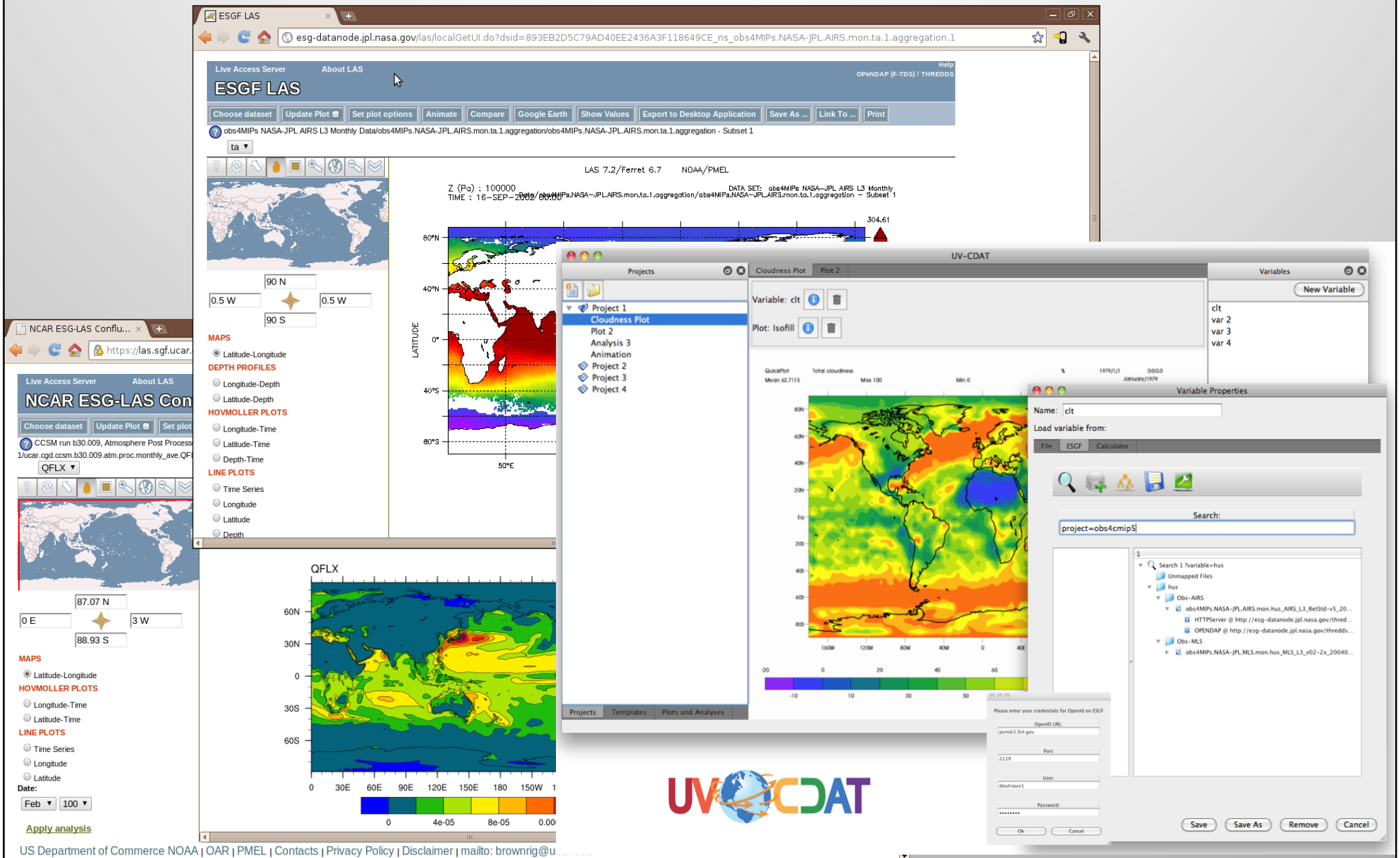


GridFTP (Globus Online and BDM) for data replication and download

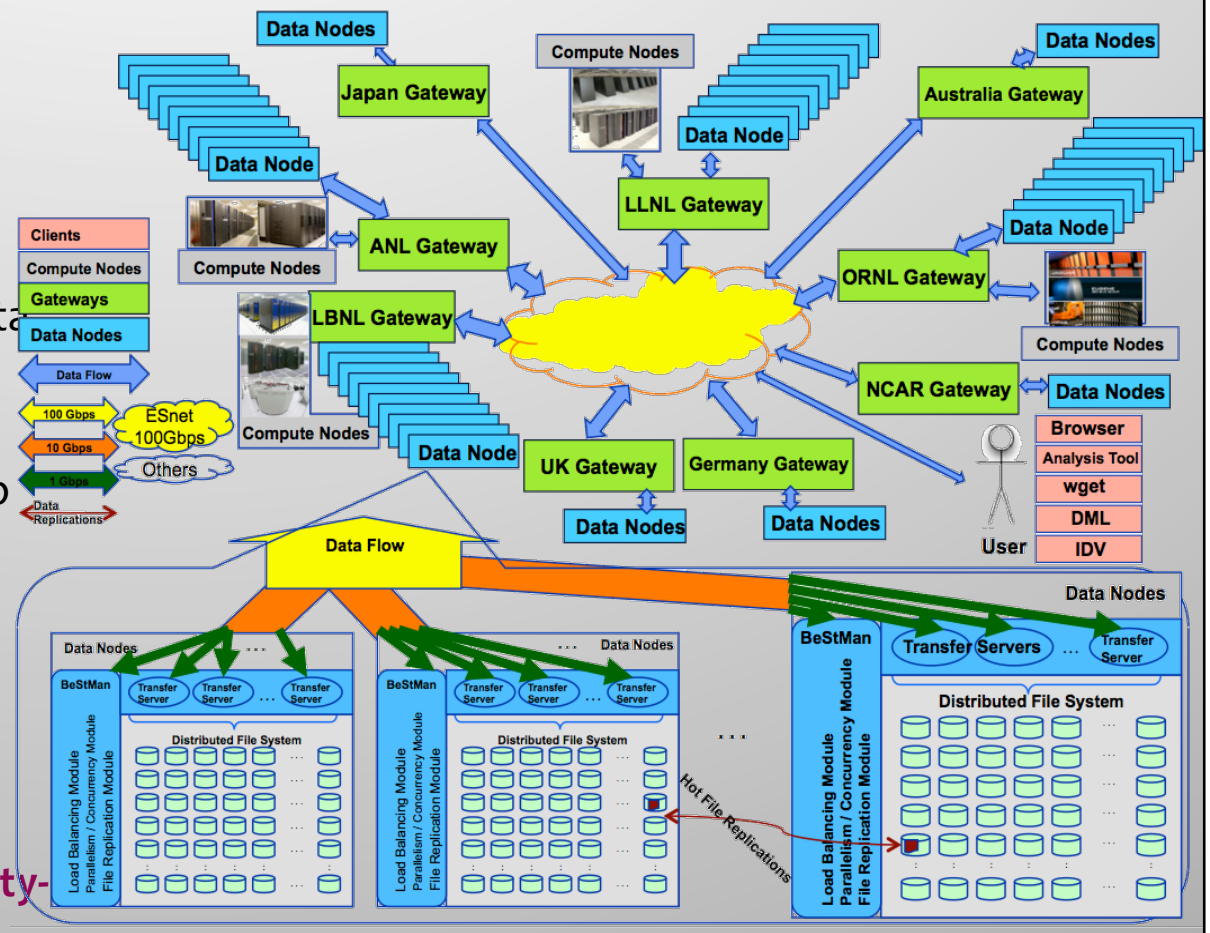


- GridFTP and Globus Online
- Berkeley Storage Manager (BeStMan)
- DataMover-Lite (DML)
- Bulk Data Mover (BDM)
- Wget

Data Preview (server-side and client-side analysis)



Examining future developments for the ESGF has potential impact on other scientific domains

- **Other scientific domains:** biology, chemistry, nuclear energy, high-energy physics, etc.
- **Peer-to-Peer Architecture – building an ecosystem**
- **Analysis** services for:
  - extremely large data sets
  - multiple large data sets are not co-located
  - cloud computing
- **Data integration** and advanced metadata capabilities
- **Advanced product services** via multiple scripting languages
- Integration of security assertion Markup Languages (SAML) identity providers
- **Measuring replication and data access patterns** in extreme scale ESG
- **Workflow and provenance**
- **Virtual Organization** management as **Software as a Service**
- **Advanced networks** as easy-to-use community resources
- Management of **open source, community-driven** software development

http://esg-pcmdi.llnl.gov/progress-reports