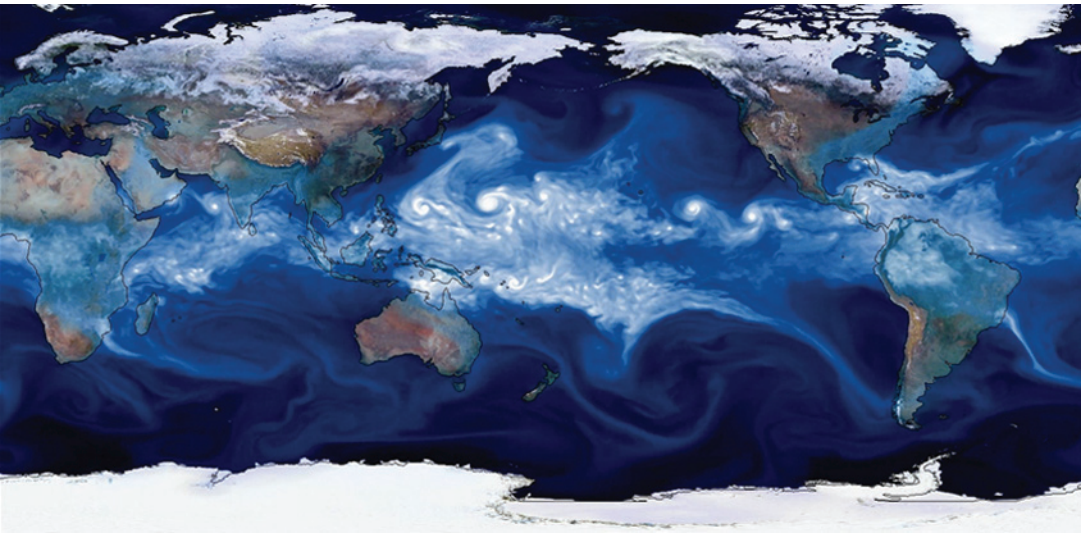
Biological and Environmental Research Network Requirements

BER Network Requirements Review
Final Report

Conducted November 29-30, 2012

ESnet
Energy Sciences Network



BERKELEY LAB
Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY
Office of Science

DISCLAIMER

Biological and Environmental Research Network Requirements

Office of Biological and Environmental Research, DOE Office of Science
Energy Sciences Network
Gaithersburg, Maryland — November 29 - 30, 2012

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR). Vince Dattoria is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract

DE-AC02-05CH11231.

This work was supported by the Directors of the Office of Science, Office of Advanced Scientific Computing Research, Facilities Division, and the Office of Biological and Environmental Research.

This is LBNL report LBNL-6395E

Participants and Contributors

V. Balaji, Princeton (ESGF)

Tom Boden, ORNL (ARM, NGE, CDIAC)

Dave Cowley, PNNL (EMSL)

Eli Dart, ESnet (Networking)

Vince Dattoria, DOE/SC/ASCR (ESnet Program Manager)

Narayan Desai, ANL (KBase)

Rob Egan, LBNL (JGI)

Ian Foster, ANL (ESGF, Data management, Globus Online)

Robin Goldstone, LLNL (PCMDI, ESGF)

Susan Gregurick, DOE/SC/BER (Biological Systems Science Division)

John Houghton, DOE/SC/BER (BER Program)

Cesar Izaurralde, PNNL (GLBRC)

Bill Johnston, ESnet (Networking)

Renu Joseph, DOE/SC/BER (Climate and Environmental Sciences Division)

Kerstin Kleese-van Dam, PNNL (Data management)

Mary Lipton, PNNL (Pan-Omics facility)

Inder Monga, ESnet (Networking)

Matt Pritchard, BADC (ESGF, CEDA, BADC)

Lauren Rotman, ESnet (Partnerships and Outreach)

Gary Strand, NCAR (ESGF, UCAR-CA)

Cory Stuart, ANL (ARM)

Tatiana Tatusova, NIH (NCBI)

Brian Tierney, ESnet (Networking)

Brian Thomas, University of California at Berkeley (Banfield Lab)

Dean Williams, LLNL (PCMDI, ESGF)

Jason Zurawski, Internet2 (Networking)

Editors

Eli Dart, ESnet — dart@es.net

Brian Tierney, ESnet — bltierney@es.net

Table of Contents

1 Executive Summary

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the U.S. Department of Energy (DOE) Office of Science (SC), the single largest supporter of basic research in the physical sciences in the United States. In support of SC programs, ESnet regularly updates and refreshes its understanding of the networking requirements of the instruments, facilities, scientists, and science programs that it serves. This focus has helped ESnet be a highly successful enabler of scientific discovery for over 25 years.

In November 2012, ESnet and the Office of Biological and Environmental Research (BER) of the DOE SC organized a review to characterize the networking requirements of the programs funded by the BER program office.

Several key findings resulted from the review. Among them:

1. The scale of data sets available to science collaborations continues to increase exponentially. This has broad impact, both on the network and on the computational and storage systems connected to the network.
2. Many science collaborations require assistance to cope with the systems and network engineering challenges inherent in managing the rapid growth in data scale.
3. Several science domains operate distributed facilities that rely on high-performance networking for success. Key examples illustrated in this report include the Earth System Grid Federation (ESGF) and the Systems Biology Knowledgebase (KBase).

This report expands on these points, and addresses others as well. The report contains a findings section as well as the text of the case studies discussed at the review.

2 Findings

2.1 General Findings

The scale of climate data sets continues to increase for both model and observational data. The growth of these data sets is expected to continue for the foreseeable future as the capabilities of high-performance computing facilities and observational instruments continue to increase. The aggregate data volume available to climate scientists is expected to reach exabyte scale within 10 years.

The Earth System Grid Federation (ESGF) now serves as a foundation for a significant amount of work done in climate science. An example of this is the Climate Science for a Sustainable Energy Future (CSSEF) project, which relies on ESGF infrastructure for data exchange. In addition, a surge in demand for data served by the ESGF is expected after the publication of the Intergovernmental Panel on Climate Change (IPCC) *Fifth Assessment Report (AR5*) in 2013 and 2014.

Several ESGF sites are experiencing data transfer performance problems, and could benefit from performance-measurement and -monitoring services such as those provided by PERformance Service Oriented Network monitoring Architecture (perfSONAR) measurement hosts. Also, some ESGF sites have been reluctant to deploy high-performance tools such as GridFTP and Globus Online due to lack of systems engineering resources, or the prioritization of other work above system performance engineering.

In some circumstances, the physical transport of portable media is still the only viable option for data transfer. One such example is observation data collected by instruments in very remote locations (e.g., some Atmospheric Radiation Measurement [ARM] program sites). In these cases, satellite connectivity is the only option, but it is very slow and very expensive. ARM (and other facilities/projects) and ESnet should periodically evaluate these conditions and determine whether improvements can be made.

Several opportunities exist for interagency collaboration in the Earth sciences area. Established through the Earth System Grid Federation (ESGF) peer-to-peer enterprise, U.S. interagency collaboration exists between DOE, the National Aeronautics and Space Administration [NASA], the National Oceanic and Atmospheric Administration [NOAA], and the National Science Foundation [NSF]. In addition, ESGF promotes international collaborations with European and Asian countries, and Australia. However, in order to facilitate greater scientific insights in climate science, more national and international collaboration opportunities must be established to meet the growing demands of extreme scale computing (i.e., storage, analysis, and visualization) on increasingly complex hardware and network systems.

KBase has two use models for ESnet. One involves user access to KBase resources hosted at KBase sites, e.g., uploading data to KBase, using KBase, and downloading results from KBase. The other involves the use of ESnet as a high-performance interconnect that enables KBase API calls and the movement of data among KBase

resources for analysis, data replication, and so forth. KBase will probably want to implement this high-performance interconnect using dedicated, long-lived virtual circuits between KBase sites. It may make use of dynamic circuits as well.

The definition of a model is needed, with supporting documentation, for the architecture of "the other end" of a genomics raw data submission to KBase, the Joint Genome Institute (JGI), or other genomics resources. Ideally, the documentation would cover systems configuration, software stack, network configuration, etc.

Bioinformatics data sets are growing at rates faster than Moore's law (5X/year for the past several years). This is a challenge for computing, networking, and storage. In some cases, it is not yet clear when the raw data associated with an experiment can be deleted. Deletion of the raw data would be desirable in many cases, as the derived data sets are typically much smaller. However, the field is evolving so rapidly that many users are keeping as much data as they can. In addition, some scientists are working on analysis algorithms for the processing of raw data — these groups need access to the raw data, regardless of whether other collaborations can work only with reduced data sets or not. This indicates a need to transfer raw bioinformatics data sets in at least some cases.

A lengthy discussion took place over whether the National Center for Biotechnology Information (NCBI) needs to preserve the raw data associated with genome sequencing. Some attendees believed that someone needs to keep the data, especially in certain cases (e.g., where obtaining another sample to sequence would be difficult in the future, or where getting the appropriate expert to annotate another sequence would be difficult). It was clear from the discussion that no good framework is currently in place for deciding which raw data sets to keep and which to delete to conserve resources.

Metadata and provenance were mentioned several times at the meeting. It appears that there would be significant scientific leverage if one collaboration were able to use data collected by another collaboration. This would only be possible if the data were annotated correctly. Currently, the incentive structures for scientific collaborations (e.g., tasks that contribute to career advancement) do not promote good annotation of data. Some changes are coming (such as the ability to cite data using digital object identifiers) that may incentivize the creation and publication of valuable data sets.

Some gene sequencer vendors have experienced pushback from customers regarding data growth rates (5X/year). It is possible that some vendors might reduce the growth rate to as little as 2X/year; however, it should be noted that this is still higher than Moore's law.

Some data movement still occurs via physical transport of portable media.

3 Action Items

Several action items for ESnet came out of this review. These include:

- ESnet will conduct a webinar covering the Science DMZ and Globus Online for members of the ARM collaboration.
- ESnet will use its contacts to help ARM with connectivity to remote sensor locations.
- ESnet will work with bioinformatics collaborations (e.g., JGI, KBase, and others) to build a networking and systems-engineering group for bioinformatics.
- ESnet will engage with the ESGF community on several tasks, including:
    - A perfSONAR test mesh and dashboard
    - Network performance tuning for ESGF data transfer nodes
- ESnet will work with KBase to document an appropriate configuration for loading raw data into KBase for analysis. This effort will start with JGI as a test case.
- ESnet will explore collaboration with KBase engineers on OpenFlow technologies.
- ESnet will engage JGI and assist with data transfer performance between JGI and the University of California (UC) at Davis.
- The Great Lakes Bioenergy Research Center (GLBRC) needs assistance with data transfers to/from Oak Ridge National Laboratory (ORNL).
- Pacific Northwest National Laboratory (PNNL) asked for assistance with setting up an On-Demand Secure Circuits and Advance Reservation System (OSCARS) circuit between PNNL and the Joint Global Change Research Institute (JGCRI) at the University of Maryland.
- The Evergreen system at the University of Maryland needs help with data access.
- ESnet will continue to develop and update the fasterdata.es.net site as a resource for the community.
- ESnet will continue to assist sites with perfSONAR deployments and with network and system performance tuning.

In addition, ESnet will continue development and deployment of the ESnet OSCARS to support virtual circuit services on the ESnet network.

4 Review Background and Structure

The strategic approach of Advanced Scientific Computing Research (ASCR) (ESnet is funded by the ASCR Facilities Division) and ESnet for defining and accomplishing ESnet's mission covers three areas:

1. Working with the DOE SC-funded science community to identify the networking implications of instruments and supercomputers, and the evolving process of how science is done
2. Developing an approach to building a network environment that will enable the distributed aspects of SC science and continuously reassess and update the approach as new requirements become clear
3. Continuing to anticipate future network capabilities to meet future science requirements with an active program of R&D and advanced development

For point (1), the requirements of the SC science programs are determined by:

(a) A review of the plans and processes of the major stakeholders, including the data characteristics of scientific instruments and facilities, to investigate what data will be generated by instruments and supercomputers coming online over the next 5-10 years. The future process of science must also be examined: How and where will the new data be analyzed and used? How will the process of doing science change over the next 5-10 years?

(b) Observing current and historical network traffic patterns to determine how trends in network patterns predict future network needs.

The primary mechanism to accomplish (a) is through SC Network Requirements Reviews, which are organized by ASCR in collaboration with the SC Program Offices. SC conducts two requirements reviews per year, in a cycle that assesses requirements for each of the six program offices every three years.The review reports are published at http://www.es.net/requirements/.

The other role of the requirements reviews is to ensure that ESnet and ASCR have a common understanding of the issues that face ESnet and the solutions that it undertakes.

In November 2012, ESnet and the BER organized a review to characterize the networking requirements of BER-funded science programs, with an emphasis on high-performance computing facilities. Participants were asked to codify their requirements in a case-study format that included a network-centric narrative describing the science; instruments and facilities currently used or anticipated for future programs; the network services needed; and how the network is used. Participants considered three timescales in their case studies: the near term (immediately and up to two years in the future), the medium term (two to five years in the future), and the long term (greater than five years in the future). The information in each narrative was distilled into a summary table, with rows for each timescale and columns for network bandwidth and services requirements. The case-study documents are included in this report.

5 Office of Biological and Environmental Research

5.1 BER Program Overview

The Biological and Environmental Research (BER) program supports fundamental research and scientific user facilities to address diverse and critical global challenges. The program seeks to understand how genomic information is translated to functional capabilities, enabling more confident redesign of microbes and plants for sustainable biofuel production, improved carbon storage, or contaminant bioremediation. BER research advances understanding of the roles of Earth's physical and biogeochemical systems (the atmosphere, land, oceans, sea ice, and subsurface) in determining climate so we can predict climate decades or centuries into the future — information needed to plan for future energy and resource needs. Solutions to these challenges are driven by a foundation of scientific knowledge and inquiry in atmospheric chemistry and physics, ecology, biology, and biogeochemistry.

BER research uncovers nature's secrets from the diversity of microbes and plants to understand how biological systems work, how they interact with one another, and how they can be manipulated to harness their processes and products. By starting with the potential encoded by organisms' genomes, BER-funded scientists seek to define the principles that guide the translation of the genetic code into functional proteins and the metabolic/regulatory networks underlying the systems biology of plants and microbes as they respond to and modify their environments. BER integrates discovery- and hypothesis-driven science, technology development, and foundational genomics research into predictive models of biological function for DOE mission solutions.

BER plays a unique and vital role in supporting research on atmospheric processes; terrestrial ecosystem processes; subsurface biogeochemical processes involved in nutrient cycling, radionuclide fate and transport, and water cycling; climate change and environmental modeling; and analysis of impacts and interdependencies of climatic change with energy production and use. These investments are coordinated to advance an earth system predictive capability, involving community models open to active participation of the research community. For more than two decades, BER has taken a leadership role to advance an understanding of the physics and dynamics governing clouds, aerosols, and atmospheric greenhouse gases, as these represent the more significant weaknesses of climate prediction systems. BER also supports multidisciplinary climate and environmental research to advance experimental and modeling capabilities necessary to describe the role of the individual (terrestrial, cryospheric, oceanic, and atmospheric) component and system tipping points that may drive sudden change. In tight coordination with its research agenda, BER supports three major national user facilities: the Atmospheric Radiation Measurement (ARM) Program's Climate Research Facility, the Joint Genome Institute (JGI), and the Environmental Molecular Sciences Laboratory (EMSL). Significant investments are provided to community database and model diagnostic systems to support research efforts.

5.2 BER Climate and Environmental Science Overview

The Climate and Environmental Sciences Division (CESD) focuses on fundamental research that advances a robust, predictive understanding of Earth's climate and environmental systems and informs the development of sustainable solutions to the nation's energy and environmental challenges. As provided by the 2012 CESD *Strategic Plan* (http://science.energy.gov/~/media/ber/pdf/CESD-StratPlan-2012.pdf), five goals frame the division's programs and investments: (1) Synthesize new process knowledge and innovative computational methods that advance next-generation, integrated models of the human-Earth system; (2) develop, test, and simulate process-level understanding of atmospheric systems and terrestrial ecosystems, extending from bedrock to the top of the vegetative canopy; (3) advance fundamental understanding of coupled biogeochemical processes in complex subsurface environments to enable systems-level prediction and control; (4) enhance the unique capabilities and impacts of the ARM and EMSL scientific user facilities and other BER community resources to advance the frontiers of climate and environmental science; and (5) identify and address science gaps that limit translation of CESD fundamental science into solutions for DOE's most pressing energy and environmental challenges.

CESD focuses on three research activities, each containing one or more programs and/or linkages to national user facilities. These activities are: (a) the Atmospheric System Research activity, which seeks to understand the physics, chemistry, and dynamics governing clouds, aerosols, and precipitation interactions, with a goal to advance the predictive understanding of the climate system; (2) the Environmental System Science activity, which seeks to advance a robust, predictive understanding of terrestrial surface and subsurface ecosystems, within a domain that extends from the bedrock to the top of the vegetated canopy and from molecular to global scales; and 3) the Climate and Earth System Modeling activity, which seeks to develop high-fidelity community models representing earth system and climate system variabilities and change, with a significant focus on the response of systems to natural and anthropogenic forcing.

The primary programs that actively use ESnet are: (1) the Earth System Modeling (ESM) Program, which develops advanced numerical algorithms to represent the dynamical and biogeophysical elements of the earth system and its components; (2) the Regional and Global Climate Modeling Program, which focuses on understanding the natural and anthropogenic components of regional variability and change, using simulations and diagnostic measures; (3) the EMSL facility, which provides integrated experimental and computational resources for discovery and technological innovation in the environmental molecular sciences to support the needs of DOE and the nation; and (4) the ARM facility, which provides the national and international research community unparalleled infrastructure for obtaining precise observations of key atmospheric phenomena needed for the advancement of atmospheric process understanding and climate models.

ESnet continues to be the primary network provider for data transfer for Coupled Model Intercomparison Projects (CMIPs), which in turn facilitate the analysis and synthesis for

the Intergovernmental Panel on Climate Change (IPCC). CMIPs are carried out by utilizing the multiple nodes of the Earth System Grid Federation (ESGF). In addition, numerous multi-lab projects, such as the Climate Science for a Sustainable Energy Future (CSSEF), use ESnet to support data transfer requirements involving the ESGF. As the emphasis on finer spatial resolution for climate and environmental models is combined with more detail on uncertainty quantification associated with model outputs, data transfer requirements become increasingly more important. ESnet is also the primary network provider that enables remote access to EMSL's high-performance computing (HPC) system, numerous mass spectrometry systems, and EMSL's Aurora data-storage archive. EMSL has also established interfaces with the JGI for automated downloading of data. All these developments are significantly increasing EMSL's data-storage needs and the associated need for users to access data remotely. ESnet has played and will continue to play an increasingly vital role in enabling the science for DOE climate and environmental research. As data volume increases for both climate models and the observational capabilities in user facilities, CESD expects increasing pressure to assure that the petabytes of data and model output are readily available to the user community through ESnet.

5.3 BER Biological Systems Science Overview

The Biological Systems Science Division supports a diverse portfolio of fundamental research and technology development to achieve a predictive, systems-level understanding of complex biological systems to advance DOE missions in energy and the environment. By integrating genome science with advanced computational and experimental approaches, the division seeks to gain a predictive understanding of living systems, from microbes and microbial communities to plants and other whole organisms. This foundational knowledge serves as the basis for the confident redesign of microbes and plants for sustainable biofuel production, improved carbon storage, and contaminant remediation. ESnet is the primary network provider that enables large-scale data transfer for the JGI with the National Center for Biotechnology Information (NCBI) and other key stakeholders.

Research into systems biology and the DOE Genomic Science program is aimed at identifying the foundational principles that drive biological systems. These principles govern the translation of genetic codes into integrated networks of catalytic proteins, regulatory elements, and metabolite pools underlying the functional processes of organisms. These dynamic interactions of nested subsystems ultimately determine the overall systems biology of plants, microbes, and multispecies communities. The ultimate goal of the Genomic Science program is to achieve sufficient understanding of the fundamental rules and dynamic properties of these systems to develop predictive computational models of biological systems and tools for rational biosystems design.

Genomic Science program research also brings the omics-driven tools of modern systems biology to bear on analyzing interactions between organisms that form biological communities and with their surrounding environments. Understanding the relationships between molecular-scale functional biology and ecosystem-scale

environmental processes illuminates the basic mechanisms that drive biogeochemical cycling of metals and nutrients, carbon biosequestration, and greenhouse gas emissions in terrestrial ecosystems or bioenergy landscapes.

The major objectives of the Genomic Science program are to:

1. Determine the molecular mechanisms, regulatory elements, and integrated networks needed to understand genome-scale functional properties of microbes, plants, and interactive biological communities
2. Develop -omics experimental capabilities and enabling technologies to achieve dynamic, systems-level understanding of organism and/or community function
3. Develop the knowledgebase, computational infrastructure, and modeling capabilities to advance predictive understanding and manipulation of biological systems

6 ARM Climate Research Facility

6.1 Background

The Atmospheric Radiation Measurement (ARM) Climate Research Facility is a long-term measurement facility funded by the Climate and Environmental Sciences Division (CESD) of BER in DOE that focuses on measuring:

- **Cloud properties**. Microphysics (phases of water), optical properties, and patterns of occurrence
- **Aerosol properties**. Size, chemistry, optical properties, and generation and decay pathways
- **Cloud and aerosol interactions**. Absorption of aerosols by clouds and cloud formation triggered by aerosols
- **Sunlight energy fate.** Radiative flux transfer, heating rate profiles, components of reflected and absorbed radiant energy, direct and diffuse light
- **Atmospheric state**. Profiles of temperature, water vapor, wind, and aerosols
- **Surface properties**. Surface fluxes, soil conditions

The ARM Climate Research Facility is building a "climatology" (multiyear record) of these measurements related to cloud formation, sunlight energy fate, aerosol formation/decay, and aerosol interactions with clouds. The measurements are used to improve parameters that represent these processes in global circulation models (GCMs). The GCMs are used for the prediction of future climate patterns. Parameters for cloud formation, sunlight energy fate, and aerosol interactions are thought to be the source of the largest uncertainties in these models and long-term climate forecasts.

ARM field sites are located in Oklahoma/Kansas, the Alaska North Slope, and the tropical Western Pacific (Manus and Nauru Islands and Darwin, Australia). The ARM mobile facility is currently located in Cape Cod, Massachusetts, and soon will be in Brazil. Formerly it was in India, the Azores, China, Germany, Niger, and coastal California. A second, more modular, mobile facility, designed to be ship-based if needed, is operating on a cargo ship in the Pacific and was previously in the Maldives and Colorado. Two additional sites are being developed and will be operational in 2014 in the Azores and the Alaska North Slope. As a user facility, ARM regularly has field campaigns colocated with existing sites that involve collaborations with the entire atmospheric community and their instruments. The program also has ties with the National Oceanic and Atmospheric Administration (NOAA), the National Aeronautic and Space Administration (NASA), and the European Centre for Medium-Range Weather Forecasts (ECMWF).

Field data systems are located at each of the field and mobile sites. Data systems with facility-wide functions are located at Brookhaven National Laboratory (BNL) and Oak Ridge National Laboratory (ORNL). A distinct aspect of the ARM data collection is that it is continuous and has essentially the same parameters for its entire history. Most other

studies in these aspects of atmospheric science include only short-term case studies of a few weeks or months.

The users of the ARM data and network resources include: ARM facility personnel (for initial data collection, internal transfer, processing, and storage) and the research community (for access/download/use of documentation and data). The user community is mostly located in the United States, but is also globally distributed. Users can be divided into the following categories: working at DOE facilities; not working at DOE but within the United States; located at universities; and persons from foreign countries.

For the ESnet use case, we will consider the network requirements of transferring data from the worldwide distributed instruments into the ARM Data Management Facility (DMF), the required access to data from ARM and other agencies to create new value-added products (VAPs), and the networking needs of the ARM user community.

6.2 Key Local Science Drivers

6.2.1 Instruments and Facilities

Each instrumented site has a local computer system with several terabytes of storage to handle buffering and on-site review of data. Most of the instruments (e.g., radiometers, meteorological sensors, aerosol samplers) are relatively small data producers (<10 MB/day). However, other instruments, such as 3-D scanning radars or lidars, produce about 15-80 GB/day per unit. ARM has 28 radar systems. The local network connects the site data system with data loggers and instrument computers and facilitates instrument uptime and quality and the continuous data collection.

The ARM data systems at BNL and ORNL are each involved in the processing functions that create or distribute higher-order data products, and each laboratory has many terabytes of storage to manage the data sets. Each system uses the “local” network infrastructure at its DOE laboratory. Within each data system, some use of private networks occurs to link multiple systems performing similar functions. The ARM Archive at ORNL shares the High-Performance Storage System (HPSS) mass storage system with the supercomputers at ORNL and has access to significant storage resources.

6.2.2 Process of Science

The use of local networks is dominated by monitoring of instruments, data collectors, data processing, data storage, and data distribution. Quality review of the data products and processes also uses the local networks. A variety of operations, scientific (instrument and quality experts), and systems personnel use the local network. Researchers have very limited and infrequent access to the local networks and this access is typically during field campaigns with a limited duration. The instrument mentors from the infrastructure have regular interactive use of the local networks.

6.3 Key Remote Science Drivers

6.3.1 Instruments and Facilities

Because of the globally dispersed nature of the ARM sites and the goal of a continuous record of measurements, the Internet is a critical component for accomplishing the ARM mission. Each site is minimally staffed and has off-site monitoring and maintenance of computer systems. ARM has a large, dispersed infrastructure team monitoring all aspects of data quality and system components. Each site uses a local ISP to connect to the global Internet. VPNs are implemented between each measurement site and ANL, which has a VPN to the ARM DMF at ORNL.

The ARM infrastructure at ANL provides VPN tunnels to each of the measurement facilities and supports the following services:

- Global and ARM infrastructure remote access to the measurement sites with access controls
- Secondary (hidden), recursive arm.gov DNS service accessible to measurement sites
- Scheduled and random security scans of measurement facilities
- Measurement facility syslog archive
- VOIP support among measurement facilities and limited access to U.S. POTS lines
- Measurement facility traffic monitoring (Snort and related tools)
- User-level VPN access to measurement facilities coming in the near future
- Measurement site network device configuration management

The following documents the current connection to the Internet for each measurement facility and possible bandwidth upgrades, should operations funding become available:

- ARM Southern Great Plains (SGP) site near Lamont, Oklahoma: Currently 100 Mbps through Oklahoma OneNet. ARM currently sends over 130 GB/day to the DMF over this link.
- ARM North Slope of Alaska (NSA) Barrow, Alaska, site: 3xT1 connection (via ATT satellite).
- Manus Island, Paupa New Guinea: Satellite link with Intelsat General (U.S. ground station in Riverside, California). 1152 Kbps/384 Kbps (uplink/downlink relative to site). A second ground station supports an off-site radar. The Manus ground stations and the Nauru measurement facility share the satellite bandwidth. ARM infrastructure is available to upgrade the link but cost is a limiter.
- Republic of Nauru: Shares the satellite bandwidth with Manus Island.
- Darwin, Australia, site: 6 Mbps through Telstra Corp. Potential to upgrade to 10 Mbps (symmetrical).
- ARM mobile facility: Currently located on Cape Cod using 6xT1 link.

Each measurement site implements a VPN through ANL to the DMF for hourly data and metadata transfer. The DMF at ORNL provides centralized access for first-order data evaluation by the ARM Data Management team. Hourly access to updates of data at the

DMF helps ensure optimal data quality and minimizes data gaps. The DMF supports the following ARM-wide services:

- Receives raw measurements and metadata from measurement sites
- Performs the "ingest" of raw data (i.e., converts raw data into a standardized NetCDF format for ease of use by ARM facility users)
- Hosts the ARM Data Quality processing systems
- Hosts arm.gov DNS
- Hosts science.arm.gov, which provides for ARM user logins and scientific collaboration via shared file resources, wiki collaboration, and other services
- Hosts measurement facility-wide local- and wide-area network capacity monitoring
- Hosts measurement facility-wide compute systems capacity monitoring
- Manages the reliability, timeliness, and completeness of all ARM data streams
- Receives data/metadata storage media from measurement sites (large data streams that cannot be delivered by Internet) for ingest and subsequent transfer to the ARM Archive at ORNL
- Hosts engineering services to design, develop, and evolve the ARM data system and data flow processes

The DMF receives ~150 GB/day from the sites over the Internet. Because of bandwidth limitations to most of the remote sites (e.g., Alaska and the oceanic islands), up to 400 GB/day is sent to the DMF via transportable hard drives.

At present, ARM pays for all its network links. See Table 1 for a list of carriers and costs.

Table 1. Carriers and costs

Site	Network Speed/Bandwidth	Costs per Month
SGPC1	100 Mbps	$3,000/mo
TWPC1, I10, C2	384 Kbps down/1152 Kbps up	$14,000/mo
TWPC3	6 Mbps	$4,400/mo
NSAC1	5 Mbps	$10,000/mo
AMFC1	9 Mbps	$4,600/mo
AMFC2	134 Kbps (200 MB/mo limit)	$650/mo (satellite)
AMFC3	1.5 Mbps	$7,300/mo

The DMF sends all site data and derivative products to the ARM Archive at ORNL. Because both are now located at ORNL, this is a local area transfer. This has been less than 500 GB/day but is expected to grow to 1 TB/day in the next year. Further growth will occur over the next few years as the secondary data products are implemented for the new American Recovery and Reinvestment Act (ARRA)-funded instruments.

The eXternal Data Center (XDC) at BNL manages the receipt of non-ARM instrument data of interest to the ARM user community from 11 data providers currently (see

http://www.arm.gov/xdc/xds). These include field campaigns and regular data sets produced by other groups (such as satellite data). The data volume varies but is on the order of 50 GB/day. The XDC also hosts the following ARM-wide services:

- Acquires and ingests external data relevant to ARM measurements (satellite data, climatological data from other agencies, etc.)
- Transfers of external data to ARM Archive
- Reprocesses external data with dedicated system located at ARM Archive at ORNL
- Hosts the http://www.arm.gov Web site, which is tightly integrated with http://www.archive.arm.gov and requires synchronizing databases between the Archive and XDC
- Hosts secondary arm.gov DNS domain
- Hosts ARM-wide LDAP service

ARM offers its users access to its data via two key means: access to processed data streams from its measurement facilities and Value Added Products. VAPs are data products created through the analysis and processing of existing data products into VAPs. In particular, these contain quantities of interest that are either impractical or impossible to measure directly or routinely. Physical models using ARM instrument data as inputs are implemented as VAPs and can help fill some of the unmet measurement needs of the ARM facility or improve the quality of existing measurements. In addition, when more than one measurement is available, ARM also produces "best-estimate" VAPs. Most VAPs are open-ended products to which new data are continually added. While many of the VAPs are solely derived from ARM measurements, others integrate the results of measurements from other agencies or are partially based on data products from other projects. This integration of external source data necessitates the regular transfer of data from other sites to the central ARM DMF processing site.

The Archive at ORNL maintains the long-term storage for all ARM data and distributes it to scientists who order the data. Data requests are submitted through a variety of Web-based user interfaces, including Web applications ranging from simple forms to interactive graphical displays. The current volume of stored data at the Archive is 300 TB, with 1 PB expected within three years. Data requested for downloads (via FTP) are currently 7-10 TB/month (400K-1,000K files). In addition, ARM recently started to publish its data to the Earth System Grid Federation (ESGF), allowing direct access to its data and data products for climate modelers in their major data-exchange environment. ORNL has a local ESGF node that it uses to publish the data; the data are further replicated from there to other sites for faster access.

In addition to data storage and distribution, the ARM Archive hosts the ARM Reprocessing Center. Large numbers of data files (hundreds of K) are transferred between the Archive and the Reprocessing Center on the local network. Remote access to the Reprocessing Center is enabled for ARM staff who perform applications software installation, review processing results, and control processing flow. The ORNL login server facilitates this access.

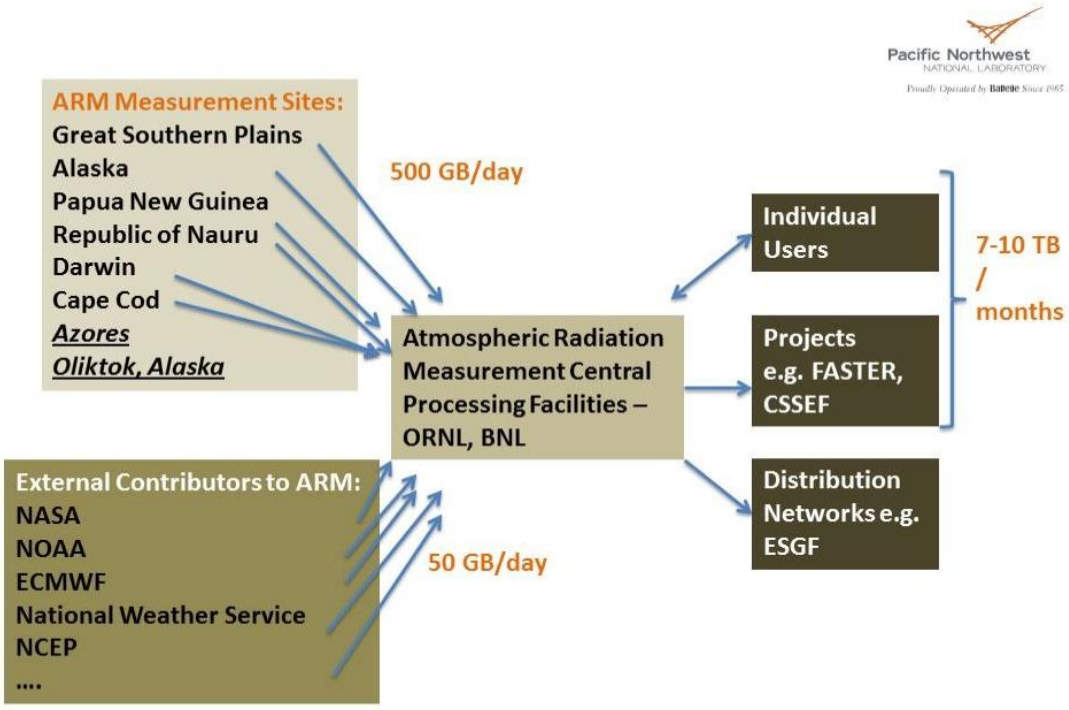


Figure 1. External network requirements for ARM.

6.3.2 Process of Science

Scientists have typically downloaded data sets relevant to their research from the Archive and performed analysis at their local institutions. This paradigm is augmented by ARM's production of VAPs that include data values based on higher-order integration, analysis, or derivation. VAPs can result in smaller data sets being downloaded. However, this may not be true from the 3-D radars. The 3-D gridded products may be larger than the original radial rays of collected data.

ARM has several developments that may reduce the data volumes to be transferred over the network. Researchers often only need a small portion of a data set for their work, and ARM is developing tools to extract and subset data in ways that make small portions of the large sets of ARM data more easily accessible to scientists. A recent survey of ARM scientists showed that a majority would have reservations about downloading more than 100 GB of data per task.

In the Climate Science for a Sustainable Energy Future (CSSEF) case study, we outline how external users and projects use Atmospheric Radiation Measurement (ARM) Climate Research Facility data.

6.4 Local Science Drivers — the Next 2-5 Years

6.4.1 Instruments and Facilities

In conjunction with the implementation of the new 3-D radars funded by ARRA, ARM is providing resources for very large data tasks (tens of TB) to enable users to perform

their work on a system adjacent to the Archive at ORNL. The implementation of large libraries of precomputed visualizations (data plots and animations) will be developed. If performance issues can be resolved, interactive visualization may also be implemented. These will increase the demand for local area data transfer.

6.4.2 Process of Science

The primary change in the scientific process associated with the much larger data volumes will be the relocation of part of the processing from the user's systems to the ARM systems. More preliminary processing for data extraction, summarization and integration, preliminary data visualization, and preliminary data analysis is likely to be performed on ARM's systems (primarily at the Archive). The primary network implications are the need for reduced latency, the secure transport of authentication processes, and the prevention of untoward activities (accidental or malicious). Many of these data-processing instructions will originate from users on foreign networks. Acquiring user specifications for data extraction via "finite" user interfaces is relatively straightforward, and constraining the results is manageable. However, specifications for more complex summarization and data-integration functions are very inefficient to formulate in a user interface and the processing requests generated outside a user interface are more difficult to constrain or review. Accommodating interactive graphics via the Internet will require limited authentication into limited command domains or proxy processes.

6.5 Remote Science Drivers — the Next 2-5 Years

6.5.1 Instruments and Facilities

In FY 2009, ARM received $60 million in capital equipment funds from ARRA to expand its instrumentation and improve its infrastructure. This funding specifically included the purchase of radar systems that generate today 10 times the amount of data than was previously being collected. In 2012, ARM received funding to deploy two more sites on the Azores and at Oliktok on the North Slope of Alaska to be complete in 2014. These sites will include a normal set of ARM instruments, including radars, and are expected to produce 200 GB/day of data. No other significant additional instrumentation is expected for the next several years.

At present, ARM has to pay for all its remote network capacity to its measurement sites, and slow transfer rates and high costs often prohibit more than the basic direct access to those sites:

- ARM (and external) radar scientists remotely interacting with the on-site radars to observe and tune scanning algorithms
- Remote monitoring of instrumentation in support of increased number of short-term field studies involving guest instrumentation

Much of its observational data has to be shipped today on hard drives to the central processing site (400 GB/day versus 150 GB/day, rising to 1 TB a day), making it impossible to get data in real time and respond to changes in data quality in a more

timely fashion. This potentially negatively influences the quality and extent of the data products ARM can deliver to its user community. With a wide variety of other data services — both in the DOE and other agencies — requiring real-time access to remote sensors, the question arises whether satellite data transfer could be integrated into the ESnet service offering, enabling the sharing of costs and bandwidth and delivering a more integrated network provision.

Thanks to the plans for on-site data analysis and visualization, we expect individual user access rates to stay level or even slightly decrease; however, the increased distribution of data via the ESGF will add to the scheduled data transfer both on site and off.

6.5.2 Process of Science

Many new, very-high-volume instruments being implemented by ARM do not have any usage history for the data products to be generated. Except for a very small number of researchers (a few tens at most), the usage of these very large data products is exploratory (working with small selections, reviewing visualizations, etc.). As the user community better understands the scientific value of these data, their vision for data products to be generated and data products to be transferred to them or analyzed by them will grow rapidly. Between more download usage and more combinations of data products (statistical or visuals), the data-flow volume on the network is likely to increase 3-5 times within the next five years. The exact growth is difficult to predict and depends largely on where the user processing is conducted (on their systems or on ARM's systems). Historically, researchers have been much more likely to use data in their processing environment. The overhead to learn the processing environment of numerous other data centers is high. Also, security issues are frequently more complex (of necessity). The next few years may continue to follow historical patterns of data use (primarily download) or may have new patterns of remote processing. The usage of very-large-scale data centers and their analytical tools is also likely to lead to more Internet-based transfer of data.

6.6 Beyond 5 Years — Future Needs and Scientific Direction

Plans for facilities beyond five years are difficult to predict for ARM. Many "never-before-operated-continuously" implementations are just beginning. The needs beyond five years will be better known after the next two to three years. It will take that long for the infrastructure and the research community to develop a next-generation view that extends beyond the very large collection of new instruments currently being implemented. A minimal and likely scenario is a continuation of the exponential increase, with a significant jump up in the trend due to the new high-data-volume field instrumentation and secondary data products.

Significant effort is being invested into finding more frequent and better ways to conduct climate research with more joint usage of observational data and simulation results. The observations and simulations will not only continue to become spatially and temporally intense, but are more likely to focus on "short-term" (decades) and regional

(subcontinental) analyses. As this occurs, intensive scanning of observations and simulations (individually and combined) to find instances of critical impacts is likely to be common. The impact assessment is more likely to access data from numerous locations with data integration and use patterns that are much less predictable.

The development of data access and analytical tools that focus on simulation results and integrated measurement observations is a final factor that will affect the future of data transfer for climate research.

6.7 Middleware Tools and Services

The program currently uses FTP for data transfer among ARM facilities. Because the program has control over the FTP endpoints, it can effectively take advantage of available bandwidth through tuning of TCP windows and other parameters and performing parallel transfers. However, at some point the program will be required to initiate transfer connections using secure authentication techniques. The actual transfer process, however, need not be secure (i.e., no need to use compute resources to encrypt/decrypt the actual data transfer).

Network-based interactions are needed that can enable remote access to users who are temporarily authenticated by program criteria to processes that have restricted “write” actions. This is needed, for example, for interactive visualization or statistical processes commonly used during initial data exploration. Procedures now in use that include distributing secure ID tokens and passwords are too slow and tedious for initial data exploration. Global solutions developed and tested by ESNet that can provide this capability would be very helpful. Allocating sufficient resources for this purpose is difficult even for projects the scale of ARM.

Scientific programs like ARM that depend on extensive use of the Internet often opt to use DOE lab-based security plans and firewalls rather than invest significant resources for independent security plans and firewalls. When the data flows for scientific programs like ARM become extreme, the performance and management of security devices like firewalls are very important. ESnet makes a valuable contribution by identifying continually improving security hardware and associated policies that can be commonly adopted across DOE for handling large volumes of internal and external data transfers. This enables the research support programs to remain focused on the particular needs of scientists. Security enclaves may provide benefits and are under discussion and limited implementation. However, implications for performance and configuration management are not yet known.

6.8 Summary Table

Key Science Drivers		Anticipated Network Requirements	
Science Instruments and Facilities	**Process of Science**	**LAN Bandwidth and Services**	**WAN Bandwidth and Services**
Near-term (0-2 years)			
• Anticipate little instrument expansion at measurement facilities after the 2 new facilities are installed in FY14	• Anticipate increased on-site field campaigns as a result of new ARRA instruments. This implies increase in remote access to both ARM and visitor instrumentation at measurement sites. • New instruments continue to increase the amount of data to be transferred. • Visualization and analysis of very large 3-D data products may be executed by scientists.	• Expanded use of 10 G networking at data centers, especially for firewalls	• Increase network capacity from remote measurement facilities to ISP as possible, given operations budget • Investigate potential for ESnet to support the ARM operation in the future to enable real-time data streaming and eliminate the need to ship hard drives
2-5 years			
• Very few additions are expected.		• Continued increasing demands for network capacity	• Anticipate increase in volume of data flow from measurement sites to DMF as bandwidth costs decrease through cheaper commercial offerings or increased provision by ESnet • The complexity of required network protocols is likely to increase.
5+ years			
• Changes are likely to continue increases for network capacity. These are very difficult to predict because of the magnitude of changes just being started.	• Intercomparison and assimilation between 3-D simulations and measurements are common.	• Continued increasing demands for network capacity	• Expanded use of very large transfers between data centers • Expanded use of analysis and visualization software control by users from remote networks

7 CEDA / BADC (on behalf of ENES)

7.1 Background

The Comprehensive Environmental Data Archive (CEDA) currently hosts four data centers: the National Centre for Atmospheric Science (NCAS) British Atmospheric Data Centre (BADC), the National Centre for Earth Observation (NCEO) Natural Environment Research Council (NERC) Earth Observation Data Centre (NEODC), the Intergovernmental Panel on Climate Change (IPCC) Data Distribution Centre, and the U.K. Solar System Data Centre, as well as small research programmes in atmospheric science and data-curation technologies. NEODC holds around 0.3 Pb of Earth Observation (EO) data and BADC holdings are at around 0.9 PB, with major expansion underway as the community makes a step change in modeling resolution.

Until recently, CEDA's IT infrastructure had grown organically as projects contributed resources to expand storage, computing, and networking capability. However, major U.K. government e-infrastructure funding investment in autumn 2011 led to the creation of two major storage, computing, and associated network resources: the Joint Analysis System Meeting Infrastructure Needs (JASMIN) and Climate and Environmental Monitoring from Space (CEMS). JASMIN has been deployed on behalf of NCAS and is a "super-data-cluster" consisting of 3.5 Pb storage co-located with analysis computing facilities. It shares infrastructure with CEMS, an equivalent 1.1 Pb facility serving the academic EO domain but partnering with commercial organizations seeking to exploit EO data for potential spin-off activities and services. JASMIN and CEMS together deploy 4.6 Pb of fast, parallel storage at the Rutherford Appleton Laboratory (RAL), connected over its own low-latency network to their own computing facilities. Satellite systems at Bristol, Leeds, and Reading Universities consist of significant disk (500, 100, and 150 Tb, respectively) coupled with additional compute resources.

The JASMIN and CEMS facilities at RAL provide a new, faster, scalable platform for CEDA to run its data centers, and also allow CEDA to offer its scientific users new resources for processing and collaboration tools co-located with the long-term archives. JASMIN includes LOTUS, a collection of virtual and bare-metal processing clusters designed not only for efficient processing of data close to the archive, but also for providing an environment for scientists to develop and test parallelized code for deployment in larger HPC contexts.

The expected exponential growth in the size of data archives, coupled with the need for scientists to work with both simulation and observation data in increasingly sophisticated ways, strengthens the need for analysis and computation to be brought to the data, and for supported and actively managed data sharing infrastructures.

7.2 Key Local Science Drivers

7.2.1 Instruments and Facilities

CEDA and hence BADC and its other data centres acts as a data repository for a wide community of scientists spanning a range of science disciplines from climate and atmospheric science through earth observation to solar-terrestrial physics.

CEDA has 10 Gbps connectivity within its local RAL infrastructure and beyond to the Thames Valley Network (TVN) and JANET onwards to the UK academic network, and to European and Intercontinental routes via a GEANT 1 Gbps connection.

The JASMIN/CEMS deployment also involved the commissioning (ongoing at time of writing) of dedicated lightpath connections from JASMIN (and hence CEDA) to U.K. supercomputing sites. The MONSooN (Met Office and NERC Supercomputing Node) supercomputer at the U.K. Met Office will make use of a dedicated 1 Gbps link to use JASMIN as "overflow" storage, while a 2 Gbps link connects JASMIN to the data store of the HECToR (High-End Computing Terascale Resource) supercomputer in Edinburgh. Additionally, 1 Gbps links connect JASMIN to the JASMIN-North satellite node at the University of Leeds, while another aims to provide dynamic lightpath connection to KNMI (Royal Netherlands Meterological Institute) or Wageningen University in the Netherlands. Additional work is ongoing to complete improvements to the "last mile" connection to CEDA services hosted on JASMIN, which should provide a significantly more efficient connection than previously achieved by CEDA services hosted on its legacy infrastructure.

7.2.1.1 List of facilities

CEDA (as facility itself)

- Petascale data archive (4 data centres)
- Compute resource (General purpose and custom analysis environments)
- Online collaborative workspace (group-based workspaces for online processing & collaboration)

CEDA as a European ESGF Data Node

As an ESGF data node, BADC replicates key data sets and hosts an ESGF gateway interface. Other European ENES institutions running ESGF include:

- IPSL (FRA)
- DKRZ (GER)
- CCMC (ITA)
- ICHEC (IRL)
- CNRM (NOR)
- CERFACS (FRA)

PRACE supercomputers

The Partnership for Advanced Computing in Europe (PRACE) is a European pool of supercomputers to which potential users can apply for access. PRACE Tier 1 partners provide resources, Tier 2 partners (e.g., United Kingdom) consume but don't provide.

UK supercomputers (not part of PRACE)

- HECToR
- 2014+ Archer (replaces HECToR)
- STFC (Science and Technology Facilities Council) / Hartree Centre
    - Blue Joule (U.K. No. 1 supercomputer: software development)
    - Blue Wonder (iDataPlex cluster)
- Met Office supercomputer
- MONSooN
- U.K. scientists also have access to U.K., PRACE, and U.S. supercomputers.

7.2.2 Process of Science

Several workflows are used:

- Accessing data for general usage
    - Climate model data
    - Atmospheric observations
    - Earth Observation data (satellite imagery & swath data)
- Running analysis against CEDA (BADC/NEODC) archive
- Running climate models
- Development of "data sharing infrastructure" technologies
    - ExArch: Distributed data processing, automating workflow of doing processing at multiple sites.

7.3 Key Remote Science Drivers

7.3.1 Instruments and Facilities

As 7.2.1 above

7.3.2 Process of Science

The activities driving CEDA-related large-scale usage network can be divided into three categories:

Category 1 : International federations

- CMIP5
- CORDEX

- ESGF (technology rather than project).
    - Many users using wget to do own transfers over non-tuned network
    - Users may not know data are coming from the United States until transfer proceeds slowly
- Climateprediction.net

Category 2: Internation bilateral projects

- WISER
    - Input data from NCAR used to generate very large simulation output in the United Kingdom
    - Unknown amounts going back to the United States (depending on quality of output)
- UPSCALE
    - UKMO & NCAS: running on supercomputer in Germany (Hermit)
        - Analysis being generated in more than one place. Ideally avoid NxN transfers but can't avoid N transfers
        - Upscale project allocated 140M core hrs / yr = 15% of daily machine capacity
        - Currently transferring 10-50 Tb / week internationally
        - 2 Tb/day arriving at CEDA (JASMIN group workspace)

Category 3 : Large data transfers

- Large transfers across tuned networks between expert sites
    - Typically using gridftp & expert knowledge
- TCRA (20th Century Reanalysis) : pulling data from NOAA for redistribution Europe-wide (actually globally since BADC can get it all online)
    - Not using ESGF technology (yet)
- ESGF Replicas
    - Bulk transfer of data to different continents to avoid per-user intercontinental transfers

7.4 Local Science Drivers — the Next 2-5 Years

7.4.1 Instruments and Facilities

The expected growth of data generated in "big data" disciplines such as Climate Science and Earth Observation, are expected to generate proportional demands on data archives and present significant challenges not only to store the data for the long term, but to enable effective use to be made of them. The JASMIN/CEMS deployment is expected to expand further and increase CEDA's capability in providing processing and analysis infrastructure alongside the data, but increasingly as part of a UK, European and International ecosystem of large-scale data sharing infrastructure components (e.g. archives, supercomputers)

Anticipated events affecting local network connectivity:

- SuperJANET upgrade due in UK over next 18 months

- Increasing UK use of [dynamic?] lightpaths for dedicated network connections to HPC facilities and specific user institutions
    - Shared network sometimes not enough. Topic/Task-specific networks in demand.
- Exploring international [dynamic?] lightpaths for large transfers.
    - GridFTP log analysis (Liu et al, REF?) of US gridftp suggests sessions large enough to justify setting up virtual circuit. Category 3 activities (large data transfers between expert sites) could usefully exploit if efficient mechanisms were in place to initiate them.
- Two major UK Strategic activities to push (a) higher resolution and (b) complexity of models. Overall, not enough supercomputing in UK to support this.

7.4.2 Process of Science

7.5 Remote Science Drivers — the Next 2-5 Years

7.5.1 Instruments and Facilities

ENES Foresight Strategy Recommendations:

1. Provide a blend of HPC facilities ranging from national machines to a world class computing facility suitable for climate applications, which, given the workload anticipated, may well have to be dedicated to climate simulations.

2. Accelerate the preparation for exascale computing, e.g. by establishing closer links to PRACE and by developing new algorithms for massively parallel many-core computing.

3. Ensure data from climate simulations are easily available and well documented, especially for the climate impacts community.

4. Build a physical network connecting national archives with transfer capacities exceeding terabits per second.

5. Strengthen the European expertise in climate science and computing to enable the long term vision to be realized.

7.5.2 Process of Science

7.6 Beyond 5 Years — Future Needs and Scientific Direction

Future plans for compute, storage, and network capabilities, any connections to any major scientific instruments that are coming in the next 5+ years

7.7 Outstanding Issues

- “Last mile” of network connections tends to be sticking point of implantation
- Policy regarding quality, output, versions etc
- Manpower vs automated tools

8 Climate Science for a Sustainable Energy Future (CSSEF)

8.1 Background

The Climate Science for a Sustainable Energy Future (CSSEF) is a collaborative project among Oak Ridge National Laboratory, Argonne National Laboratory, Brookhaven National Laboratory, Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Pacific Northwest National Laboratory, and Sandia National Laboratories, together with the National Center for Atmospheric Research to transform the climate model development and testing process and thereby accelerate the development of the Community Earth System Model's (CESM's) sixth-generation version, CESM3, scheduled to be released for predictive simulation in the 5-to-10-year time frame. Four research themes are addressed in the project:

1. A focused effort for converting observational data sets into specialized, multi-variable data sets for model testing and improvement
2. Development of model development testbeds in which model components and submodels can be rapidly prototyped and evaluated
3. Research to enhance numerical methods and computational science research focused on enabling climate models that use future computing architectures
4. Research to enhance efforts in uncertainty quantification for climate model simulations and predictions

These four themes are mutually reinforcing and tightly coupled around three overarching research directions:

1. Development, implementation, and testing of variable-resolution methodologies that enable computationally efficient simulation of the climate system at regional scales
2. Improvement in representing the hydrological cycle and quantification of the sources of certainty in its simulation
3. Reduction and quantification of uncertainties in carbon cycle and other biogeochemical feedbacks in the terrestrial ecosystem

CSSEF will be structured to first deploy expertise in the research theme areas across the development of the atmosphere, ocean, sea ice, and land surface model components, and later to the fully coupled system.

The CSSEF project's focus is to address the DOE Office of Biological and Environmental Research's long-term measure to "deliver improved scientific data and models about the potential response of the Earth's climate and terrestrial biosphere to increased greenhouse gas levels for policy makers to determine safe levels of greenhouse gases in the atmosphere."

Figure 2 provides a conceptual schematic of the complete development and application enterprise required to construct, integrate, test, and deploy climate models as complex and comprehensive as CESM. In this view, model development and application occur

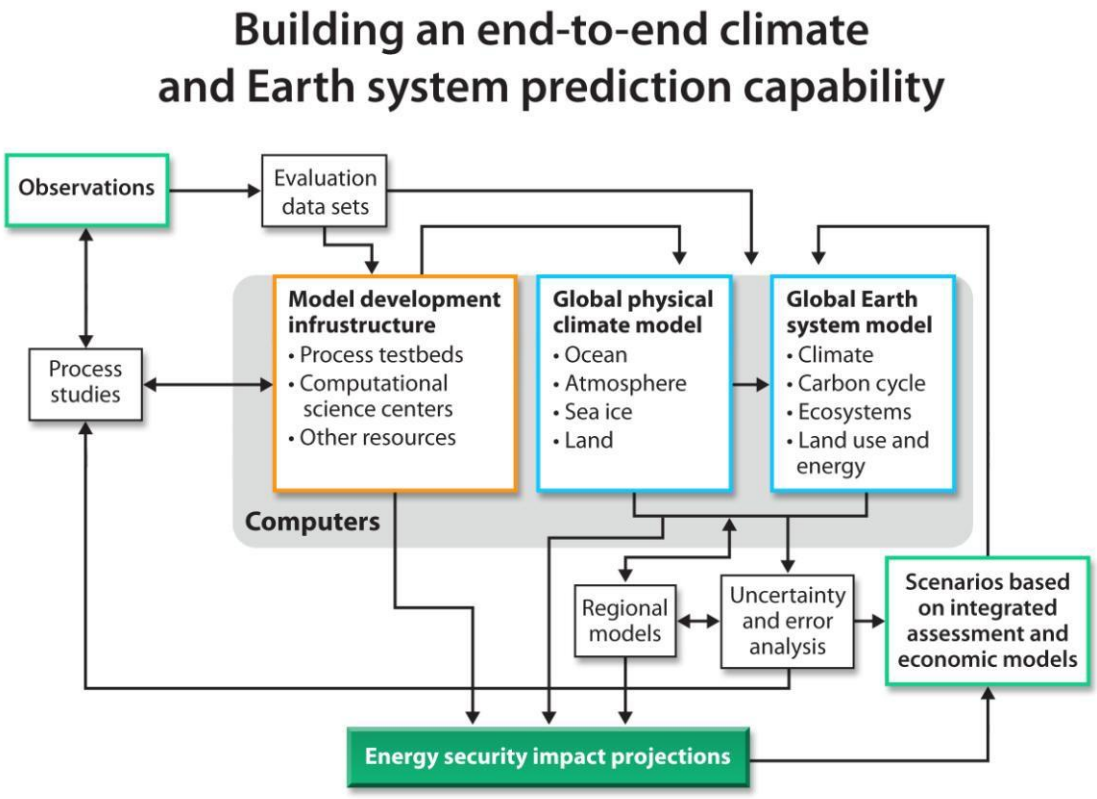


Figure 2. Conceptual view of an ongoing climate simulation and prediction enterprise such as the CESM project. New versions of models are developed from increased understanding gained through the integration of observations, process research, and earlier model studies.

simultaneously. Furthermore, simulation results continue to be analyzed and used by scientific communities for several years after the completion of the model runs. The philosophy behind CSSEF is based on the understanding that many aspects of model development and evaluation, from early research through full implementation and testing, take 10 to 15 years to accomplish, while new versions of climate models are released at approximately five- to six-year intervals for predictions, projections, and applications.

To accomplish and support the complex development efforts within the three science domains of CSSEF, the project is developing complex testbeds that facilitate information and data flow, as well as giving access to all required resources (data, storage, computing) when needed. Figure 3 presents a schematic outline of these testbeds in their final stage.

In the testbed, users can access these capabilities through clients such as Ultrascale Visualization-Climate Data Analysis Tools (UV-CDAT), which integrate resource access and usage to compute, data, and storage facilities.

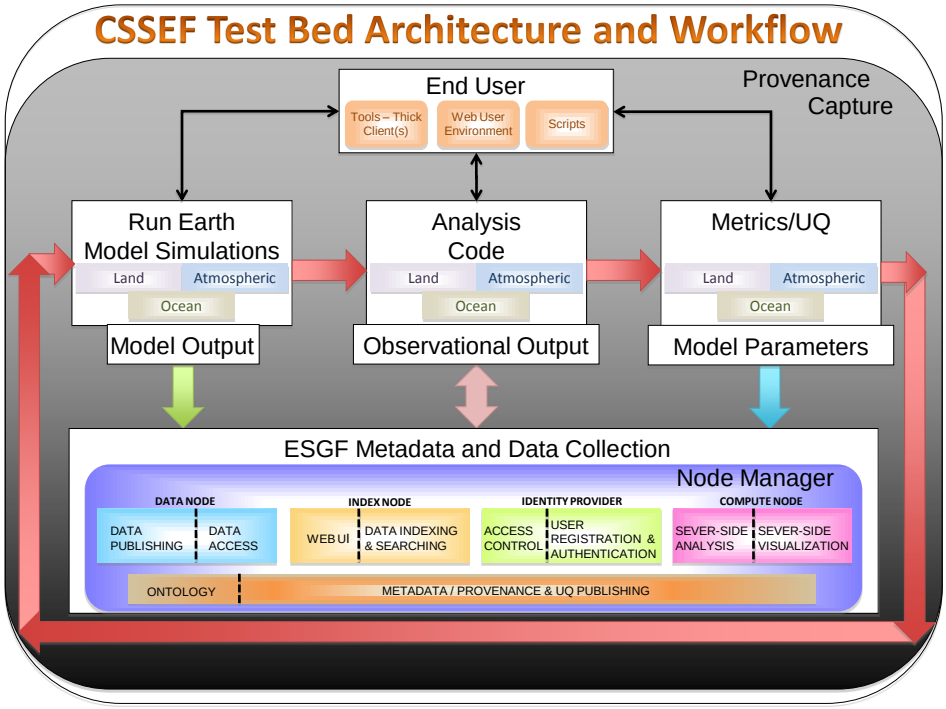


Figure 3. Conceptual view of the climate model testbeds to be developed by CSSEF

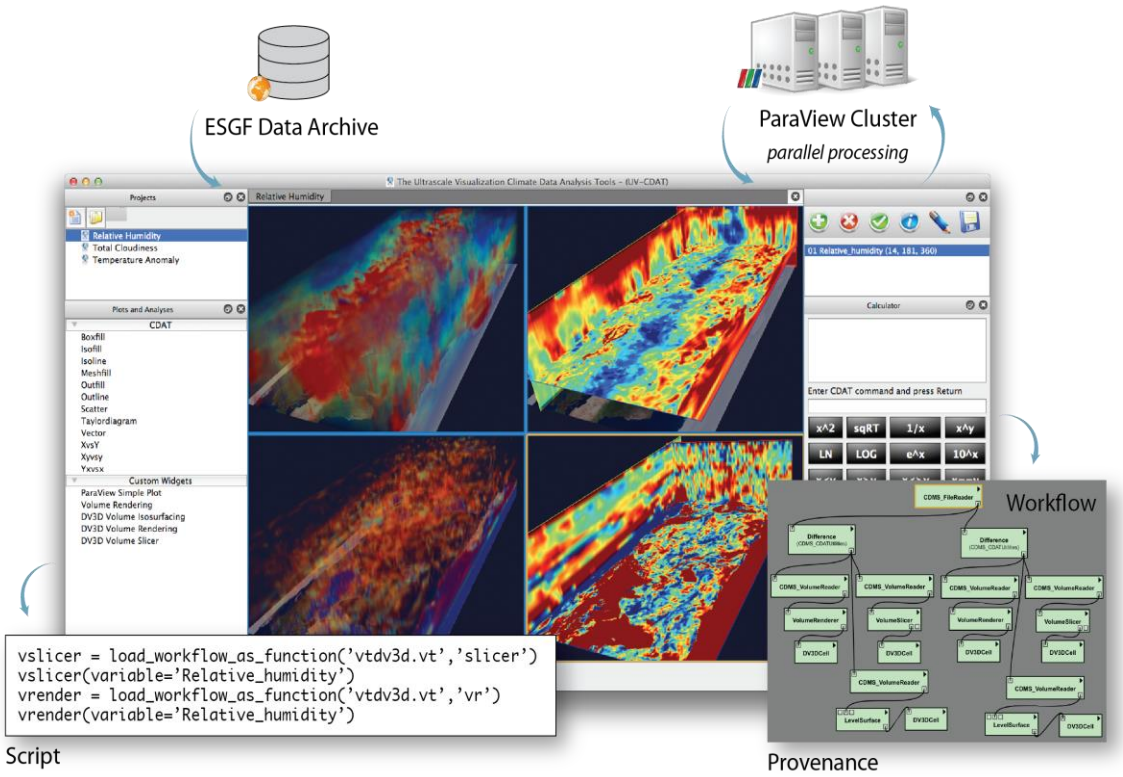


Figure 4. CSSEF environment showing UV-CDAT accessing the CSSEF ESGF Data Archive

8.2 Key Remote Science Drivers

CSSEF depends in most of its work on remote resources and so we will focus on those and their network requirements, rather than on the local networking needs of individual scientists.

8.2.1 Instruments and Facilities

CSSEF does not have any facilities or instruments of its own, but instead makes heavy use of existing DOE facilities and services. Core to its operation is access to data services for atmospheric, land, and ocean data; the Earth System Grid Federation (ESGF) for data exchange; analysis resources through tools such as UV-CDAT and the Exploratory Data analysis ENvironment (EDEN); and large computing facilities for their modeling work. Data transfers will be increasingly carried out using Globus Online. In addition, CSSEF needs access to more modest computing facilities to process and transform observational and modeling data into the right form. It relies very much on the collaboration of experts from different domains and organizations to accomplish its goals (data experts, modelers, and uncertainty quantification [UQ] specialists, as well as support from computer science); the ESGF plays hereby a central role in facilitating the data exchange among the project partners. However, because it at present does not possess integrated computing facilities, much network transfer is required to download and move data to other sites for processing or upload results.

8.2.2 Process of Science

CSSEF has three distinct science domains, with at times varying scientific workflows and networking requirements. The following describes two examples of those workflows in more detail to highlight the complexity of the interactions and resulting network requirements.

8.2.2.1 CSSEF Land Model usage

Figure 5 below describes the workflow use cases and requirements for the CSSEF Community Land Model (CLM) Uncertainty Quantification effort, particularly in support of parameter sensitivity analysis efforts.

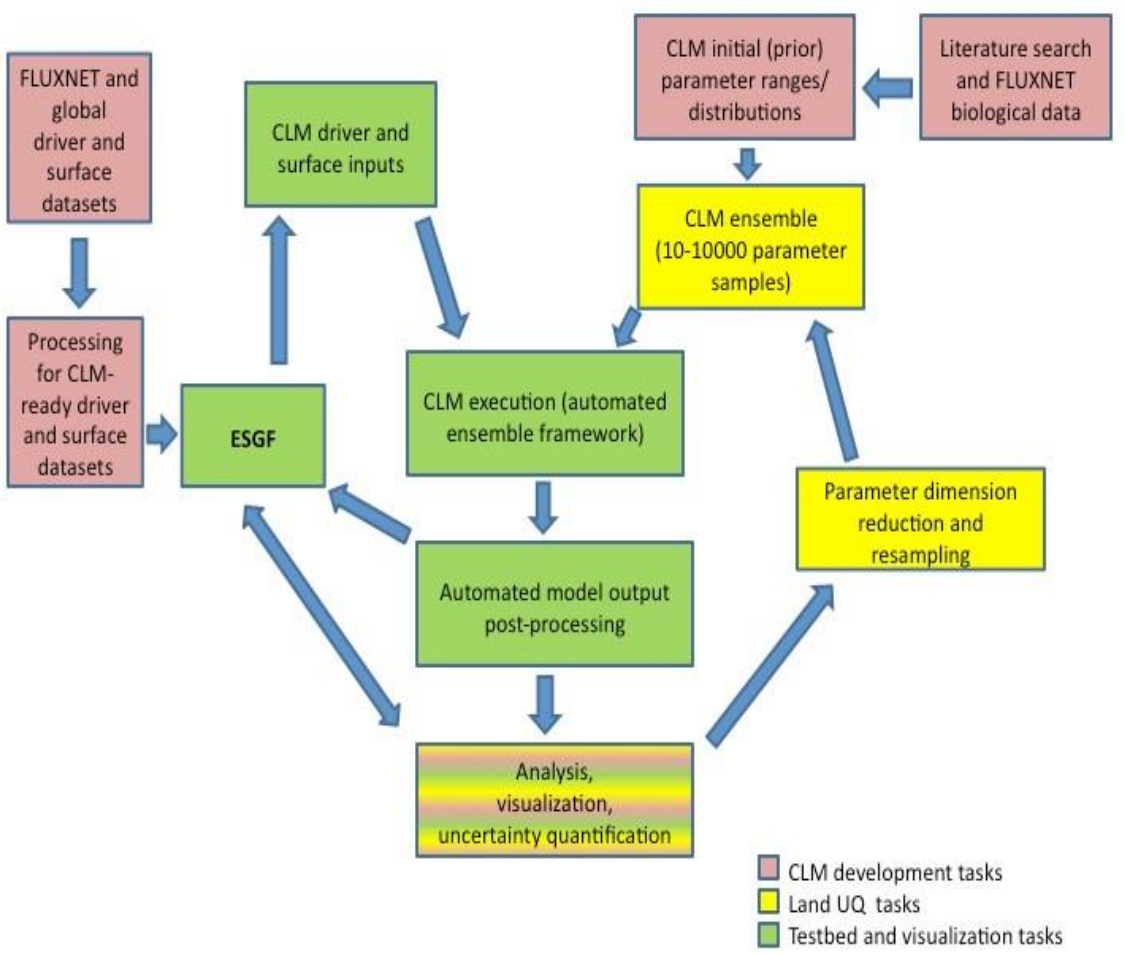


Figure 5. Conceptual overview of the data flows. Each box in the diagram is discussed below. Each arrow in the diagram represents the need for network transfer of data, the majority of which is via ESnet.

Data Experts **retrieve FLUXNET and global driver and surface data sets.** The upper-left box represents observational data that has been collected as well as future data that has not yet been collected. Typically, eddy flux and meteorological observations are made hourly and arrive in the form of "Level 4" quality-controlled data from the Carbon Dioxide Information Analysis Center (CDIAC) for U.S. sites and from FLUXNET for global sites. Ancillary biological data arrive in a standardized spreadsheet format. Further processing of these data sets is necessary for use by CLM (see below).

Processing at CSSEF site for CLM-ready driver and surface data sets. This box, which appears just below the box described above, represents further processing that should be done on this data before it is published in ESGF, such as data-quality checks, gap filling, and perhaps temporal scaling if needed. Tools for doing the needed tasks have already been developed. It will be necessary to attach metadata to the data sets to track

them through the process, answering questions such as, “Is this data set ready for ingest to CLM? Is it ready to be published for access by the community?” This metadata will be used to help prevent users from doing analysis or simulation with inputs that are incomplete or improperly vetted.

Publish new data sets to ESGF. The Earth System Grid Federation[2] is a federation of sites committed to long-term storage of high-value climate data to make it available to the scientific community for analysis and simulation. The ESGF will be used as a repository for storing data sets to be used as inputs to simulation runs, as well as simulation outputs to be shared among the project partners and wider climate community. To the extent it makes sense, publication of data to and retrieval of data from ESGF will be automated to reduce the likelihood of human error and the amount of manual work to be done.

Climate Modelers **download CLM driver and surface inputs.** The preprocessed CLM-ready driver and surface data sets as described above will be retrieved as necessary from ESGF as part of an automated framework to launch an ensemble of CLM runs.

Literature search and FLUXNET biological data. Study of the literature and currently available biological data will be used to establish initial estimates for ranges for each of the model parameters of interest.

CLM initial (prior) parameter ranges/distributions. The results of the literature search above will be standardized into a file containing the ranges for each model parameter. This parameter range file will be used by the CLM ensemble code to generate a series of samples for each parameter.

Execution of CLM ensemble runs (10-10,000 parameter samples) at Leadership Computing Facilities (LCFs). Based on the initial parameter range estimates, a number of ensemble runs will be generated to examine model behavior given various combinations of parameters. Each parameter will be varied by a small amount to assess the sensitivity of the model to variations in that parameter.

Managing such large numbers of runs is expected to be beyond the capacity of a person. This task will require an automated framework to monitor which parameter combinations have been tried and which have not, track the progress of each job to its completion, generate well-structured directory trees to organize the output files, and so forth.

Automated model output post-processing and publication of data to ESGF. Once a job completes, a determination must be made as to which output information to retain. This process is expected to generate far more data than can feasibly be kept long term. Data to be retained will be pushed into ESGF. Again, managing the quantity of data and number of files to be deleted or pushed to ESGF will be beyond the capacity of a human and will need to be automated based on rules defined in the testbed.

Aspects of this step are not well understood, and the work done in this step is expected to evolve over time. For example, it seems likely that metadata will need to be added to

the output at this point. It further seems likely that the list of metadata elements to be added will change as the project gains experience with how the output data are discovered, retrieved, and used. One likely kind of metadata that has been suggested is probability distribution functions for specific parameters.

UQ experts **retrieve ensemble results for analysis, visualization, uncertainty quantification.** Data will be retrieved from ESGF for analysis and visualization, or may come directly from job post-processing. Some analyses may be fed back to ESGF to share with other researchers or as a starting point for future studies. Uncertainty quantification analysis will be used to reduce parameter dimensions for further narrowing of the viable parameter range. This step in the process is not expected to lend itself readily to automation, as we don't know what to expect at this point.

UQ Experts and Climate Modelers **define new parameter dimension reduction and resampling.** Based on the outputs of the various simulations, the parameter ranges will be refined to further explore and define the model's sensitivity to parameter variations, thus quantifying the model's uncertainty.

Restart workflow potentially with amended data sets or directly with new ensemble runs.

8.2.2.2 CSSEF Atmospheric Model Evaluation

The following use case describes the processes associated with analyzing data sets produced by the Community Atmosphere Model (CAM). The use case described below includes descriptions of analysis performed on observational data sets, allowing for comparisons between analysis products generated from either model or observational information.

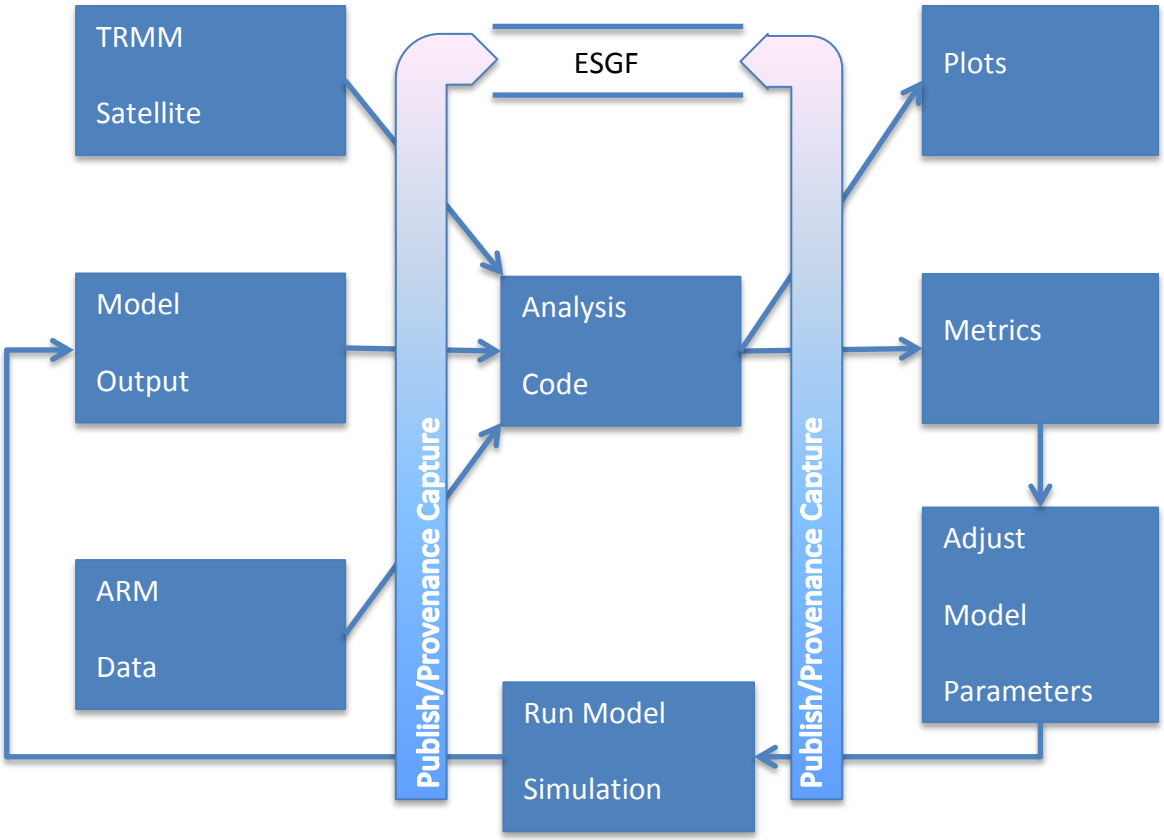


Figure 6. Overview of the data flows supporting a CAM analysis process. All data products (TRMM, ARM, model results, plots, and metrics) are published to and retrieved from the ESGF data store. Provenance information is captured from data flows and stored in ESGF as well. Simulations are run on LCF systems and data processing and analysis is carried out on local systems. Each arrow represents a data transfer stream, mostly via ESnet. The elements are described below.

Data Experts **download and transform Tropical Rainfall Measuring Mission (TRMM) satellite data on local resources.** This box represents the TRMM satellite data sets. TRMM was selected as an example of a source for satellite data; there are many others that could be used to support CAM analysis. TRMM is a joint mission between NASA and the Japan Aerospace Exploration Agency (JAXA) designed to monitor and study tropical rainfall. In order to be run using the National Center for Atmospheric Research (NCAR)

analysis code, the satellite data must have comparable variables, temporal resolution, and spatial resolution to the output of a CAM model. Or, a transformation must occur to (1) imitate the variables present in the CAM model output and (2) project the data into the temporal and spatial resolution used by the CAM model output.

Data Experts download model output from ESGF. This box represents model results/output generated by a CAM model run. The model output consists of a set of history files for each model run:

- Files of monthly means for each variable in the model output
- Files of mean or instantaneous values of model output variables at varying temporal scales (hourly, daily, every model time step)
- Files of aggregated variables on a yearly (ANN) or seasonal basis (DJF, MAM, JJA, SON)
- Files of specific variables and/or across specific geographical regions collected across the full model run

The length of each model run can range from a few minutes to hundreds of years, so the size of the data output is very variable. The length of each model run has a direct effect on what analysis can be performed on it. Output from UQ and sensitivity runs is similar to output from normal CAM runs except there will be an output data set for each iteration of the UQ run.

Data Experts download and transform ARM data. This box represents the Atmospheric Radiation Measurement (ARM) observational data sets. These data sets are representative of the general case of ground-based observational data and the set of derived data products from these measurements. ARM was selected as an example of ground-based observational data; there are others that could be used to support CAM analysis. Similar to the TRMM satellite data box, the ARM data must have comparable variables, temporal resolution, and spatial resolution to the output of a CAM model. Provenance information equivalent to that captured for the TRMM satellite data will also be captured for the ARM data.

Climate Modelers, Data and UQ Experts use analysis code to integrate and compare observational and modeling results at local sites. This box represents the analysis code selected to operate on the model output data, the satellite data, and/or the ARM data. There are many possibilities for analysis codes that can go in this step. The use case descriptions below are based around the following two example analysis codes:

1. The team at NCAR (Brian Medeiros and Rich Neale) has developed an initial set of analysis scripts using the NCAR Command Language (NCL) to calculate information about the precipitation from CAM model results.
2. Code to analyze the output of a set of UQ runs: The description here will be to generalize the types of analysis performed on UQ runs and does not refer to a specific set of codes.

It should be noted that a wide range of existing and potentially new analysis codes can and will be used to analyze the output of model runs, satellite data, and ARM or other

observational data. The software framework that surrounds this piece must be flexible, adaptable, and able to evolve as analysis needs evolve.

The NCAR analysis scripts operate on a single model run or observational set. However, they do produce comparable outputs so that in downstream analysis the scientist can compare these analysis outputs from multiple runs or output from a model run against an output from satellite data (See *Plots* and *Metrics* for descriptions of the output products created by these scripts).

In general, the UQ analysis code takes the outputs over all model cases and produces analysis about the behavior of specific variables across the full set of UQ model runs. Statistical distributions of the behavior of specific output variables are produced. Also produced are response surfaces that represent how output variables change as changes are made to multimodel parameters.

Plots. This box represents plots created by the Analysis Code step. The NCAR analysis code produces plots. Each plot is linked to a particular model output run and depends on the specified variable, temporal scale, and region used when running the analysis script. The NCAR scripts were initially set up with precipitation as the specified variable.

UQ analysis codes may produce the following plots:

1. Plots of probability density functions (pdfs) of output variables across all model runs
2. Plots representing response curves

Metrics. This box represents metrics created by the Analysis Code step. The NCAR analysis code produces the following output data products. Again, the initial example setup is precipitation.

1. Pdf of precipitation
2. Composite diurnal cycle for every grid point in the specified region
3. Calculates the amplitude, phase, and variance associated with the diurnal cycle

These same analysis products can be created for TRMM satellite data and ARM observational data as well.

UQ analysis code produces:

1. Pdfs of output variables across all model runs
2. Multidimensional response curves that represent how the model output variables change with changes to the set of model parameters

An additional set of analysis metrics can be created by using the results from this initial set of analysis metrics. As an example, if comparing a model run with satellite data, these plots and metrics could be created for a variety of interesting variables. Metrics could then be calculated as to how close the model run agreed with the satellite data (the results of which are fed into Adjust Model Parameters).

Adjust Model Parameters. Based on analysis of results from the Analysis Code step and the metrics and plots produced, model parameters may be adjusted to attempt to better answer the scientific question being asked. The CAM model is then set up and re-

launched with the new parameters. It would be beneficial to capture the analysis inferences and thought process that go into deciding the hypothesis that it will test in the next step.

Some examples are given below:

1. **NCAR scripts.** A comparison is made of the set of plots made from a model output run against a set of plots made from satellite data looking at a specific geographical region. It is determined that the model is producing too much rainfall in the region. An exploration of other model output variables is made (potentially including comparisons with a model control run) and hypothesis developed as to which model parameters should be adjusted to correct the amount of rainfall. The CAM model is set up and launched with a new set of parameters designed to test this hypothesis.
2. **UQ analysis.** After analysis of the pdfs of output variables, several model parameters that were varied in the experiment are determined to have a significant effect on model output. A new experiment is designed to further quantify the effect of these variables by running an additional set of model runs with more samples taken (and possibly samples taken at finer granularity) for these model parameters. The parameter set for this new experiment is created and the CAM runs are initiated.

Climate Modelers, Data and UQ Experts will upload any interesting or relevant results to ESGF to share with their collaborators and potentially the wider community.

Climate Modelers **run model simulation on LCF.** This task can encompass several possible goals for running the CAM model. A non-inclusive list is given below:

1. A control model run of a tuned model for comparison against model run with changes or varying parameters
2. A model run of a new model version to test the parameter set
3. A model run to test a new version of a particular model subfunction (e.g., new physics code, new parameterization code)
4. Model runs to compare with observational data
5. A set of model runs varying a defined set of parameter values to test model sensitivity to those parameter values
6. A set of model runs varying a defined set of parameters to quantify uncertainty in the model
7. Others

ESGF. This box represents an ESGF data node. The ESGF data node is used to store published data sets, making them available to the scientific community. This includes observational data sets, model input and output data sets, analysis products (e.g., plots, metrics), as well as other content deemed to have high value for the community. ESGF will provide authorization mechanisms to limit or expand access to a published data set. The ESGF data node is also responsible for storing provenance information.

8.2.2.3 CSSEF network requirements

CSSEF uses a range of facilities managed by others but because its testbeds are not yet fully automated, it lacks the infrastructure to measure the amount of network traffic it creates. It does, however, have a number of indicative figures that show the potential usage once the system is fully functional. One of its key measures is the data sets that have been published into ESGF, at present only a semi-automated process; strongly rising figures are expected once Easy-Pub (user-driven automated data publishing and sharing) software has been released to CSSEF scientists. Since the start of CSSEF in June 2011, the project has published over 160 TB. Table 2 lists the specific data sets that have been published.

Table 2. CSSEF specific data sets published between June 2011 and September 2012, listed by publishing ESGF node

Institution	Data Set	P	Type	Description	Use	Status	Size
ANL	HOMME	H	Model	Gridded	Atmosphere	Published	94 GB
ANL [Holding data for PNNL]	CAM5 UQ [F19_F19]	H	Model	Gridded	Atmosphere	Published	5.6 TB
ANL [Holding data for PNNL]	CAM5 UQ [256 x 16]	H	Model	Gridded	Atmosphere	Published	620 GB
ANL						Total Published	6.3 TB
LLNL	CAM5 CESM1 – CAM5	H	Model	Gridded	Atmosphere	Published	5.4 TB
LLNL	CAM4	H	Model	Gridded	Atmosphere	Published	166 MB
LLNL	CAM5 YOTC	H	Model	Gridded	Atmosphere	Published	472 MB
LLNL	CAM5 UQ [F19_F19]	H	Model	Gridded	Atmosphere	Published	1.2 TB
LLNL	CAM5 UQ [nond01]	H	Model	Gridded	Atmosphere	Published	13 TB
LLNL						Total Published	20.2 TB
ORNL	CESM Ultra High Res (T85)	H	Model	Gridded	Atmosphere Land Ocean	Published	48 TB
ORNL	CESM Ultra High Res (T341)	H	Model	Gridded	Atmosphere Land Ocean	Published	88 TB
ORNL	ARM CMBE	H	Model	Single Point Gridded	Atmosphere	Published	11 GB
ORNL	C-LAMP	H	Model	Single Point Gridded	Land	Published	146 GB
ORNL	AmeriFlux L2	H	Observational	Gap-filled surface weather forcing data (U.S. sites)	Land	Published	830 MB
ORNL	CDIAC	H	Observational	Gap-filled surface weather forcing data	Land	Published	132 MB
ORNL						Total Published	136.2 TB
PNNL						Total Published	0
SNL						Total Published	0
						Total CSSEF Federated Archive	162.7 TB

The following tables, using the example of atmospheric modeling in CSSEF, show the observational data sources to which users are expected to require access during their work (there are similarly complex lists for the land and ocean domains). To date, only a few have been accessed and distributed via ESGF. Each of these data sets is unlikely to be used as is, but will be further analyzed and transformed to include parameters that can be simulated by the modeling codes and are at the right scale and aggregation level. This manipulation of the retrieved data might also include the integration and synthesis of data from different products and data providers.

In addition to observational data, the modeling teams will produce tremendous amounts of data themselves, in particular for their UQ combinatorial runs. One such example is the CAMUQ data set published on ESGF at 5.6 TB for one small use case.

While the current early base data acquisition and production rate is around 15 TB/month, the scientific process still requires the data to be moved among different partners and sites, leading to a larger transfer bandwidth requirement than the basic figures suggest. The current multiplier is estimated to be 2-3, leading to up to 45 TB that needs to be transferred per month.

Table 3. Sample metrics for calibration platform and data sources

Metric	Data Source
Surface T, RH, PRECIP, CF, and radiation (time series, diurnal cycle, probability density functions)	Climate Modeling Best Estimate (Xie et al. 2010)
Higher-order and 2-D precipitation statistics (drizzle frequency, rain area, stratiform/convective partitioning)	Scanning precipitation radars at SGP, Darwin, and Manus sites
CFADs of cloud radar reflectivity and Doppler velocity to evaluate cloud vertical structure and microphysics	ACRF cloud radars + instrument simulators modified for ground-based instruments (Haynes et al. 2007; Bodas-Salcedo et al. 2008, Chepfer et al. 2008, Fan et al. 2009)
Diabatic heating (time series, diurnal cycle, lead-lag correlation of diabatic heating profile relative to surface precipitation time series)	ACRF multiyear variational analysis at SGP and Darwin, and radar-derived latent heating (Xie et al. 2004; Schumacher et al. 2004)
EIS, LTS, CAPE (Klein and Hartmann, 1993; Wood and Bretherton 2006)	ACRF radiosonde observations at all sites; Raman lidar and/or AERI retrievals at SGP and ACRF tropical sites
Relationships between variables: water vapor vs. precipitation; regime decomposition such as stratification by vertical velocity (Holloway and Neelin 2010; Bony et al. 2004)	Various ACRF measurements, re-analysis data for vertical velocity

Table 4. Sample diagnostics of validation platform and their data sources

Diagnostic	Data Source
Hurricane statistics	IBTrACs (Knapp et al. 2010)
Global precipitation and tropical precipitation intensity statistics	CMAP (Xie and Arkin 1997); GPCP (Huffman et al. 1997); TRMM (Kummerow et al. 2000)
Cloud/precipitation vertical structure statistics (satellite simulator CFADs)	Calipso/CloudSat instrument simulations (Bodas-Salcedo et al. 2008)
Drizzle Incidence in boundary layer clouds	CloudSat (Berg et al. 2010)
Multivariate relationships between water vapor, precipitation, and radiative forcing	SSM/I (Peters and Neelin 2006); Reanalysis data (Wheeler and Kiladis 1999); ERBE/HIRS (Bennhold and Sherwood 2008)
Standard core diagnostics: large-scale wind, temperature, and humidity; cloud fraction; SW and LW radiation	ECMWF or other reanalysis (Kalnay et al. 1996, Uppala et al. 2005); ISCCP (Rossow and Schiffer 1991); CERES (Wielicki et al. 1996)
MJO diagnostics	MJO working group (Kim et al. 2009)
Water cycle components (evapotransporation, water vapor transport, etc.) on continental and ocean basin scales	NEWS team (NEWS 2007)

8.3 Remote Science Drivers — the Next 2-5 Years

8.3.1 Instruments and Facilities

CSSEF will continue to use external instruments and facilities; however, it is expected that network requirements will grow as the testbeds mature, facilitating much more automated data access and creation patterns.

8.3.2 Process of Science

The project is currently in the process of establishing a range of initial testbeds for model development. As time progresses, these will be increasingly automated and serve a wider set of use cases and specific scientific enquiries, therefore strongly increasing the need for streamlined data access, processing, modeling, and evaluation and thus data transfer between resources.

8.4 Network and Data Architecture

CSSEF differs from many other projects of its type in that it aims to use existing facilities and infrastructures rather than develop its own distributed model. Consequently, it is able where possible to benefit from fast network connections and reduced data transfer needs. Figure 7 below highlights the benefit of this decision, focusing on the example of the use of ARM data within the CSSEF context. The figure shows data movement from initial collection to usage.

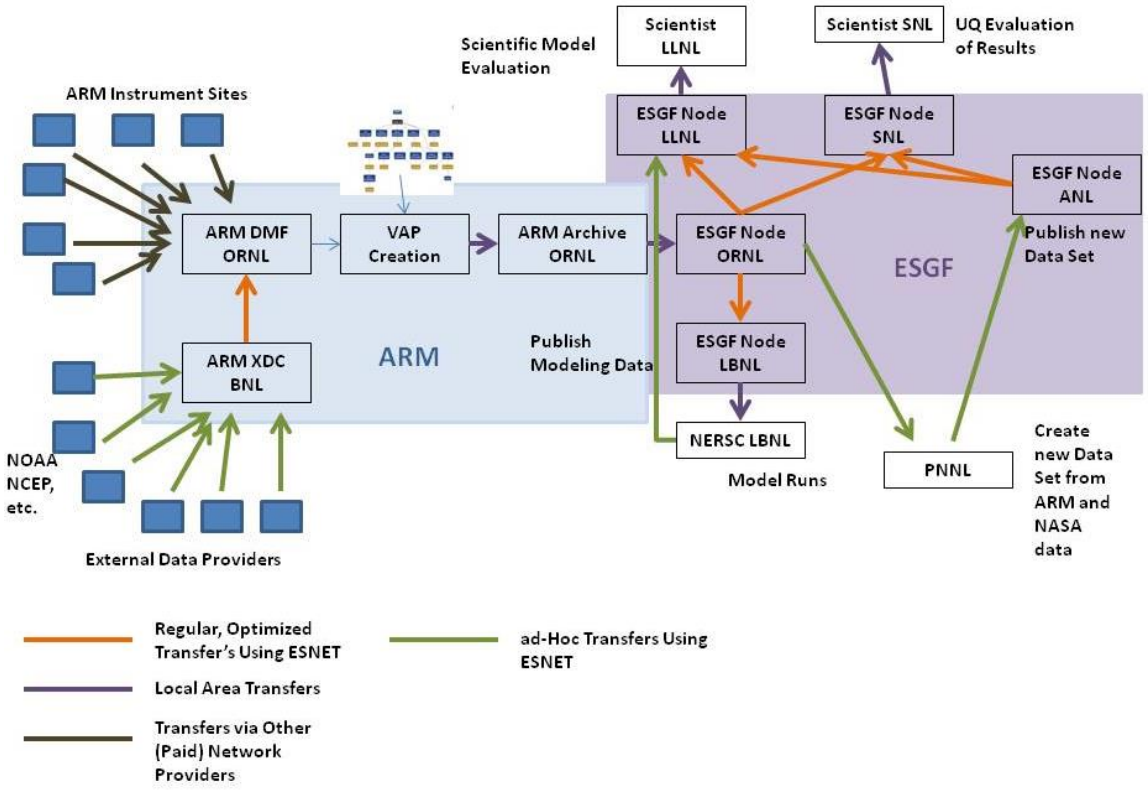


Figure 7. Data flow and network usage between ARM and ESGF in the context of the CSSEF project — focused on observational atmospheric data

The remaining ad hoc data transfer is caused by the need to manipulate and use the observational data. Users with on-site access to an ESGF data node will have less ad hoc traffic than users at sites such as Pacific Northwest National Laboratory (PNNL) that currently don't have an ESGF node. Offering capabilities at data-processing sites such as ARM or within the ESGF to process and utilize data directly could further minimize the need for ad hoc transfers. In addition, if new data sets could be contributed back to the ARM site, and —after application of quality control — redistributed to the wider community, the need for others to carry out the same data processing (frequently done by the different user groups) would be reduced. Without the integration of ARM and ESGF, the data flow would look quite different, including mostly ad hoc, unmanaged, nonoptimized data transfers that often cost more time or come in greater volume than individual scientists can accomodate.

The ARM data-flow example highlights the benefit of a hub (ESGF) providing access to all data needed by a collaborative project. It reduces the need for ad hoc data transfers between individual scientists, makes versioning possible, and helps reduce duplication of work. Furthermore, results of interest can be shared with the wider community without further data movement. Unfortunately, few observational data providers are publishing into ESGF and not all DOE laboratories and their collaborators have local access to an ESGF site. Further, facilities like ARM would make a valuable contribution to the community by assessing and improving data quality and creating new products.

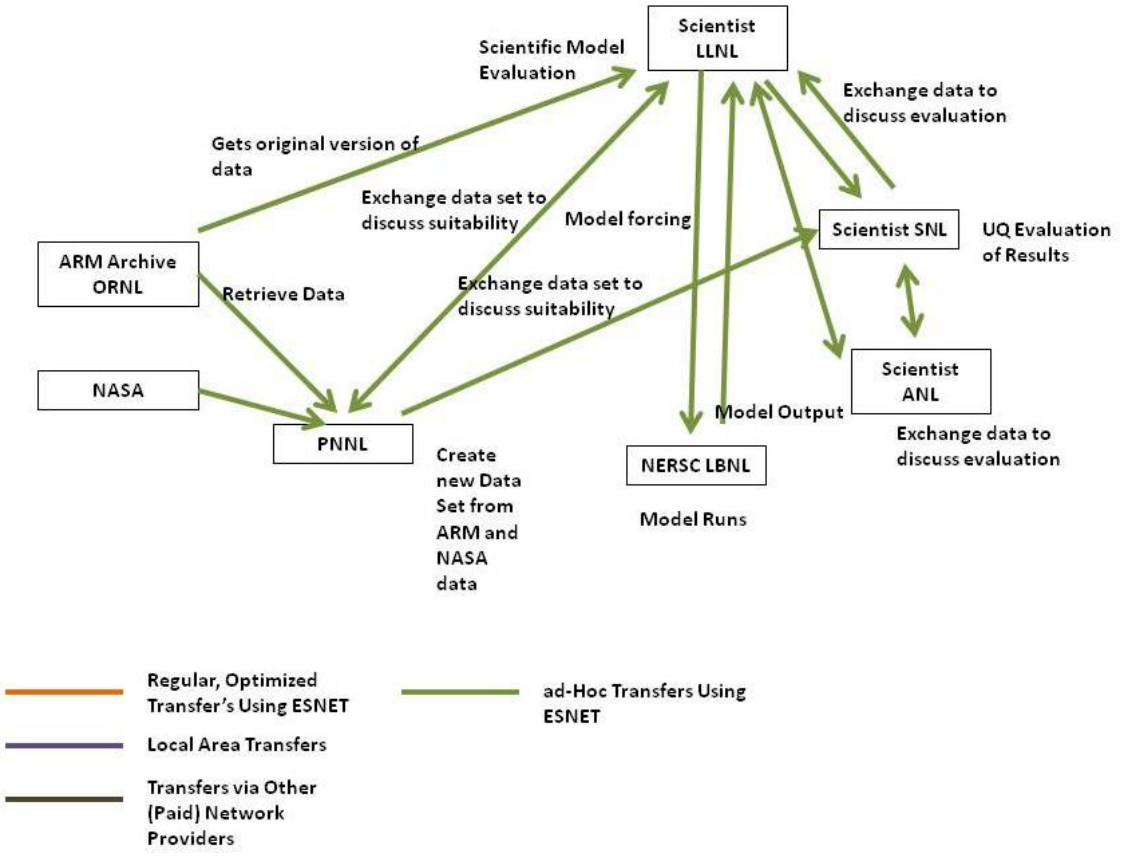


Figure 8. Project-based data transfer without support of the ARM-ESGF integration as central hub for data sharing

This integration not only as data source but also as data clearinghouse (data quality, processing, and publishing) would not only greatly improve the quality of the data available to the research community, but also reduce data transfer needs by offering easily accessible data-processing capabilities with access to most of the data and tools needed.

It is potentially beneficial to support new or renewing projects and facilities in their architectural planning by taking these available infrastructures and facilities into consideration, enabling them to support their collaboration or interaction with their community more effectively, making use of more optimized data transfer services, and reducing overall network load against a background of exponentially rising data volume.

8.5 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• New CLM and CAM simulations will continue to be added, observational data will be added to support calibration and diagnostics activities, and UQ is expected to ramp up particularly for the CLM community. Provenance Environment (ProvEn) Services will be made available to the CSSEF grid, making the origin of the data sets available to a broader community.	• Climate scientists will use CSSEF ESG-CET to store, search, and access CESM-related data sets along with observational data to perform diagnostics and calibrate. Scientists involved in UQ studies will plan, construct UQ input decks, run, and analyze their results in an iterative fashion.	• 500 TB/ month (for the entire CSSEF grid), resulting in 3.75M files	• 50 TB (75K files) in 1 hour to transfer data to high-bandwidth data transfer node	• 150 TB/week (1.1M files) • Data transferred to sites hosting data nodes at LLNL, ANL, ORNL
2-5 years				
• As ESG-CET data node codes are hardened, new nodes are expected to be deployed. Analytical services will be provided on data nodes as well as local nodes. Based on project requirements and user community demands, the number of sets is expected to increase dramatically.	• Better automating UQ studies with workflow technology • Advancing cross-referenced and provenance-related searches for grid data	• 100 TB/ week at each hosted CSSEF site • Data set composition 750K files	• No change from 0-2 years	• 150 TB/week per site (1.1M files) • Data are transferred to site ESG node, wide area needs will increase as external users access data
5+ years				
• The number of new data sources is expected to dramatically increase as well as a number of new nodes across the grid.	•	• Data volume 5 PB/ month	• 100 TB/hour, 24x7	• 10 PB/month

9 DOE-UCAR Cooperative Agreement for Climate Change Prediction Program

9.1 Background

The cooperative agreement between DOE and the University Corporation for Atmospheric Research (UCAR) sponsors the development, enhancement, use, and analysis of the National Center for Atmospheric Research (NCAR)/DOE/NSF Community Earth System Model (CESM), one of the world's most complete and advanced climate models. CESM has participation from a very large community of scientists, and peer-acceptance, which is important to ensure excellence and relevance. Because of the complexity of the problems and the technical sophistication of the models and computer codes, major modeling programs are no longer single-principal-investigator research projects. They are major technology-development efforts, and are both shared research tools and major code projects. The CESM community enables access to contributions from multiple sources in an open development process that allows incorporation and testing of a wide range of ideas in a broad spectrum of disciplines. The CESM program also has a mission to foster the creative involvement of university researchers and students in the subject area, contributing to the development of highly trained investigators. The CESM program is a complement to the other major modeling programs in the U.S. Global Change Research Program (USGCRP) that are specifically oriented toward a government mission to provide decision-support information.

The scientific objectives of the CESM program are to:

1. Develop and continuously improve a comprehensive Earth modeling system at the forefront of international efforts to understand and predict the behavior of Earth's climate.
2. Use this modeling system to investigate and understand the mechanisms that lead to interdecadal, interannual, and seasonal variability in Earth's climate.
3. Explore the history of Earth's climate through the application of versions of the CESM suitable for paleoclimate simulations.
4. Apply this modeling system to estimate the likely future of Earth's environment, in order to provide information required by governments in support of local, state, national, and international policy determination.

Under the auspices of the DOE-UCAR cooperative agreemet (CA), CESM simulations are carried out on a number of supercomputers, including the NCAR/University of Wyoming Yellowstone system, the DOE systems at Oak Ridge National Laboratory (Titan), the National Energy Research Scientific Computing Center (NERSC) Hopper system, Argonne National Laboratory's Intrepid system, and others. Additionally, CESM is utilized for large international model intercomparison projects (MIPs), including the Coupled Model Intercomparison Project Phase 5 (CMIP5), Transpose-AMIP (TAMIP), the Geo-engineering MIP (GeoMIP), Paleoclimate MIP Phase 3 (PMIP3), and similar projects.

Results from these simulations are often transferred between the various computing sites for analysis, depending on specific aspects of the simulations involved. The volume

of data transferred from one site to others can vary considerably, from a few hundred megabytes (MB) to tens or hundreds of terabytes (TB). Efforts are made to keep model results local to the system upon which they were generated, but that isn't always possible, especially in regard to MIP-related simulations. For example, a previous version of the model, CCSM3, was used for the 2004-2007 CMIP3 simulations, and all the data (totaling about 10 TB) was transferred (via 1 TB external hard drives) from NCAR to the Program for Climate Model Diagnosis and Intercomparison (PCMDI), as it was the host institution for the CMIP3 archive. The current MIPs (CMIP5, TAMIP, GeoMIP, and PMIP3) use the expanded and enhanced Earth System Grid Federation (ESGF) to distribute data. NCAR is the host site for the CESM-generated data from these simulations.

Nearly all the data from these CESM simulations are made available to the community via the ESGF, including the original model output and post-processed data.

9.2 Key Local Science Drivers

9.2.1 Instruments and Facilities

Table 5 describes the local computing and other resources the DOE-UCAR CA uses for carrying out simulations with the CESM, as of late 2012.

Table 5.

Site Name	Name and Type	Processors	Memory	Disk Storage	Archival Storage Capacity
NCAR-Wyoming Supercomputing Center (NWSC)	Yellowstone IBM iDataPlex	72,288 CPUs	115 TB	11 PB	100 PB

9.2.2 Process of Science

The typical process for the use by the DOE-UCAR CA of the CESM for knowledge discovery involves an experimental design created by either an individual scientist, small NCAR group of scientists, or one of the CESM working groups (a collection of scientists and others with a common interest). Once the design is finalized and the necessary resources (computing, storage, and so on) are determined, the project applies for those resources at the computing center. When the resources are allocated, the model is executed at the center; the output is analyzed, archived, and made available via the ESGF; and papers are written and submitted to various science journals detailing what was learned from the experiments.

Figure 9 is a schematic illustrating the general workflow and science process for the DOE-UCAR CA's use to CESM.

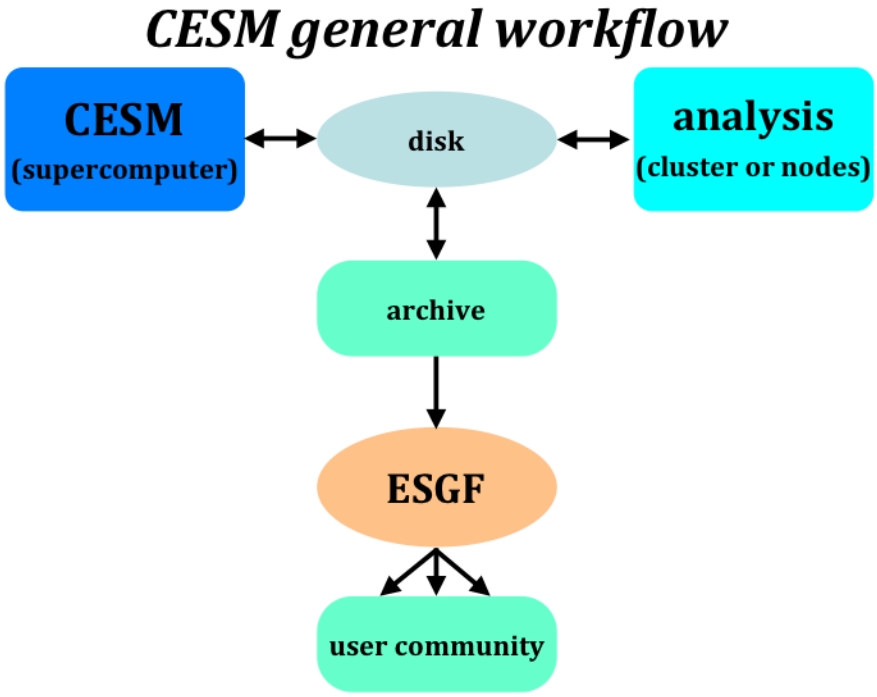


Figure 9. CESM general workflow and science process

9.3 Key Remote Science Drivers

Table 6 describes the remote computing and other resources the DOE-UCAR CA uses for carrying out simulations with the CESM, as of late 2012.

9.3.1 Instruments and Facilities

Table 6. Instruments and facilities

Site name	Name, Type	Processors	Memory	Disk storage	Archival Storage Capacity
Oak Ridge Leadership Computing Facility (OLCF)	Titan Cray XK7	299,008 CPU, 18,688 GPU	710 TB	10 PB	50 PB
National Energy Research Scientific Computing Center (NERSC)	Hopper Cray XE6	153,216 CPU	217 TB	2 PB	200 PB
Argonne Leadership Computing Facility (ALCF)	Intrepid IBM Blue Gene/P	164,000 CPU	80 TB	7.6 PB	24 TB

9.3.2 Process of Science

The same basic process for CESM simulations executed at NCAR is done at other DOE computing centers. Experiments are designed; resources allocated; and the simulations are run, post-processed, analyzed, and made available via the ESGF. Some original model output and post-processed data from these simulations is transferred back to NCAR, but because all of the DOE-UCAR CA computing resources are associated with "nodes" in the ESGF, it is not necessary to transfer all of the data just for the purpose of making them publicly available.

To convey an idea of the total data volume generated by hundreds of CESM simulations, Figure 10 uses the archival volumes at NCAR and the DOE sites to extrapolate CESM data holdings for period 2013-2018, using the 2000-2011 period for extrapolation.

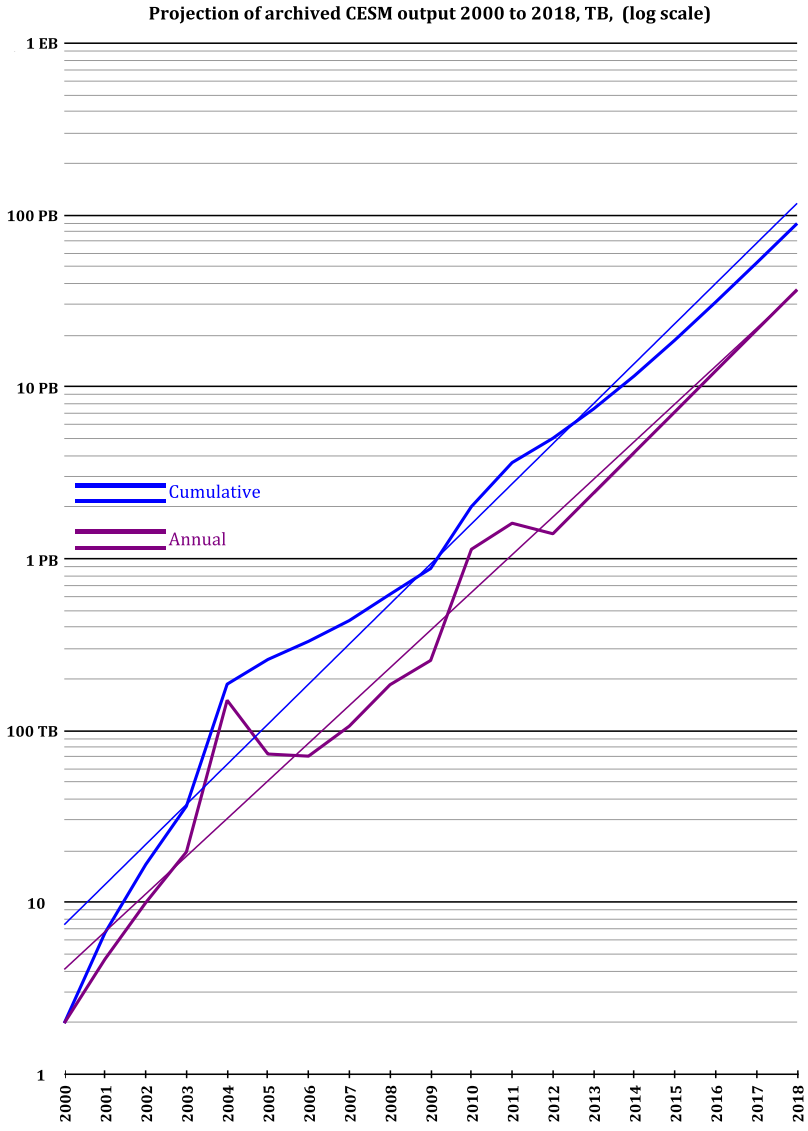


Figure 10. Projection of archived CESM output 2000-1018, in TB

9.4 Local Science Drivers — the Next 2-5 Years

9.4.1 Instruments and Facilities

Over the next two to five years, the Yellowstone supercomputer at the NWSC is expected to be upgraded in terms of processor cores, memory, disk storage, and other resources.

9.4.2 Process of Science

The science processes for the DOE-UCAR CA's use of the CESM are expected to be very similar to the current usage over the 2015-2018 time range, with the possible exception that the model output will be written in a transposed format compared with the current history format. This shift — from putting all fields from a single time period in a single file to writing all time periods for each individual model output field into a single file — will reduce the requirement to post-process the model output to make it more usable for the community.

One key project that will probably take place near the end of this period is the anticipated Coupled Model Intercomparison Project Phase 6, CMIP6. Very preliminary discussions of the scope of this project by the World Climate Research Program's (WCRP) Working Group on Climate Modeling (WGCM) have just begun; it is unknown how many model simulations the global climate modeling community will be asked to undertake. The scientific knowledge discovery process from the current CMIP5 is in its initial stages, so the possible realm of simulations requested for CMIP6 is unclear.

The current ESGF architecture may be enhanced and expanded over this period, so that any CESM CMIP6 simulations will remain resident at their host sites, without the need to transfer large volumes to model data to NCAR or between the sites.

9.5 Remote Science Drivers — the Next 2-5 Years

9.5.1 Instruments and Facilities

Over the next two to five years, it is expected that the supercomputers at each of the remote computing sites (ORNL, NERSC, and ANL) will be upgraded in terms of processor cores, memory, disk storage, and the other resources. The exact nature of the hardware available in the 2015-2018 time frame is difficult to assess, but hybrid supercomputers consisting of many CPU and GPU cores are expected.

9.5.2 Process of Science

Just as with the NWSC resource, it is anticipated that the computing resources at the DOE sites will be utilized to carry out the simulations with CESM in accordance with the DOE-UCAR CA plans, as well as the CMIP6 project.

Figure 11 illustrates the projection of archived model output from CESM for the period 2018 onward.

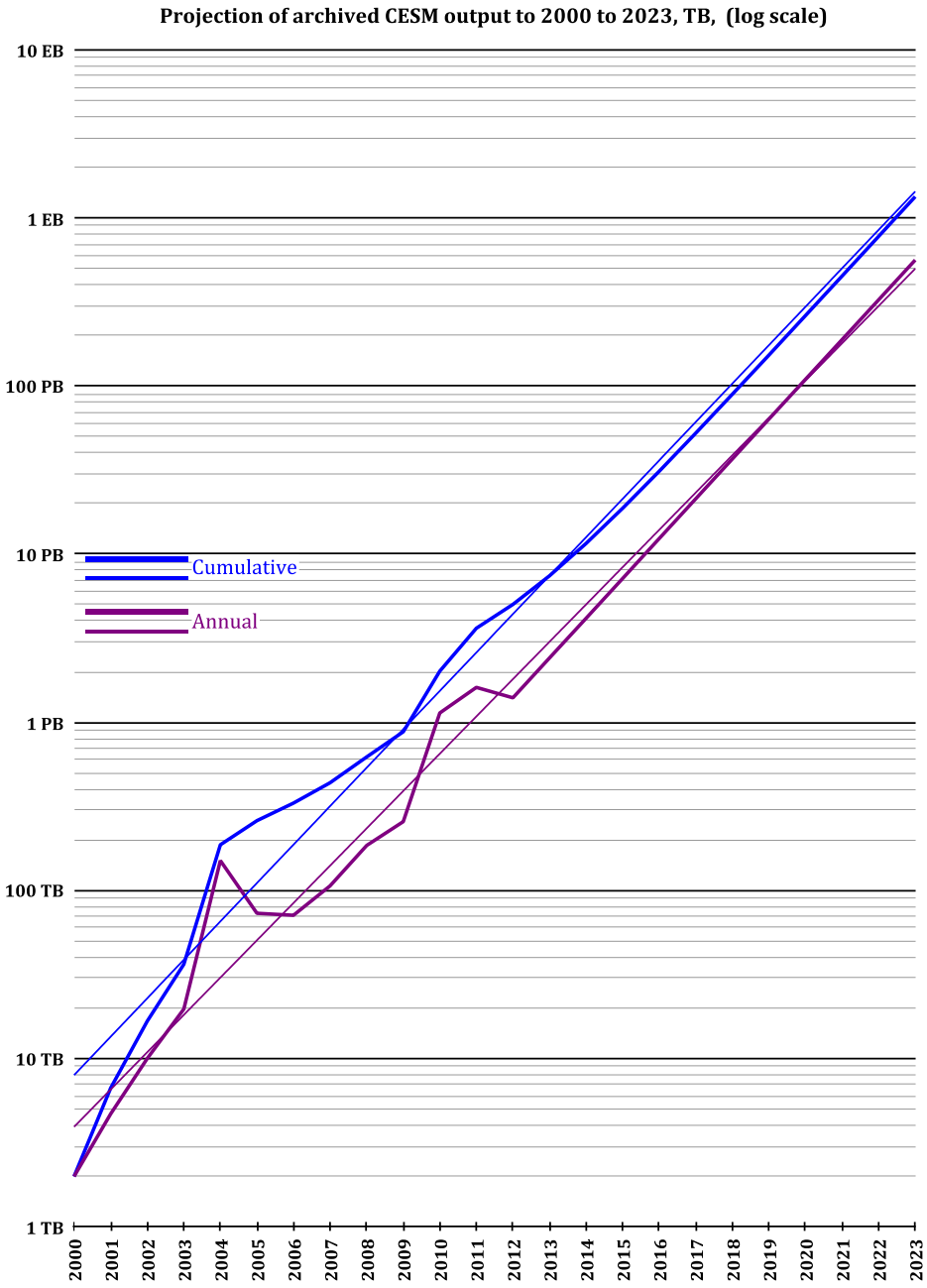


Figure 11.

9.6 Beyond 5 years — Future Needs and Scientific Direction

See *Summary Table* below.

9.7 Network and Data Architecture

The CESM project as a whole may participate in future Big Data initiatives, but has not so far. The current CESM Data Management and Data Distribution Plan is available at http://www.cesm.ucar.edu/management/docs/data.mgt.plan.2011.pdf.

9.8 Collaboration Tools

The weekly meeting of the DOE-UCAR CA team uses ReadyTalk's services for remote collaborator call-ins, as well as sharing the desktop of the meeting convener. Skype is used on occasion to collaborate with colleagues at remote locations. It's not anticipated that these practices will change.

9.9 Data, Workflow, Middleware Tools, and Services

The most significant change to current DOE-UCAR CA practices will be the likely alteration in the format of the CESM output, from the previously described history format to single-field format. This change will reduce the overall data throughflow from the model execution to disk and then archival storage, and will provide the global user community with easier, more efficient access to CESM results.

The DOE-UCAR CA and CESM will continue to rely on the ESGF and its successor projects to publish and deliver model output to the user community. Other projects may be incorporated into the ESGF to enable data format changes (to other binary formats, from netCDF to GIS-compatible formats, for example) and the ability to extract, subset, and additionally process the model results. Whatever tools ESGF makes available will be exploited by the DOE-UCAR CA.

9.10 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• NCAR-Wyoming Supercomputing Center (NWSC), • Oak Ridge Leadership Computing Facility (OLCF) • National Energy Research Scientific Computing Center • Argonne Leadership Computing Facility (ALCF)	• Simulations with CESM carried out at each individual computing center, output post-processed on site, with nearly all data made available via the Earth System Grid Federation (ESGF)	• Maximum ~8 TB/day, average ~1-2 TB/day • Average file size ~600 MB, range from ~100 MB to ~1.5 GB	• Max rate of ~300 GB/hr from supercomputer to disk; similar rate for disk to archival tape	• ~10 TB/week • Data may be transferred from ORNL, NERSC, ANL to NCAR, using Globus toolkit.
2-5 years				
• Upgrades of all supercomputing sites • CMIP6	• Simulations with CESM carried out at each individual computing center, output post-processed on-site, with nearly all data made available via ESGF	• Maximum ~100 TB/day in 2018; • Average file size ~2 GB, range from ~200 MB to ~4 GB	• Max rate of ~500 GB/hr from supercomputer to disk; similar rate for disk to archival tape	• ~5-~10 TB/day • Data may be transferred from ORNL, NERSC, ANL to NCAR, using Globus toolkit and successors.
5+ years				
• Upgrades of all supercomputing sites • CMIP7?	• Simulations with CESM carried out at each individual computing center, output post-processed on-site, with nearly all data made available via ESGF	• Maximum ~900 TB/day in ~2022 • Average file size ~6 GB, range from ~5 to ~10 GB	• ~500 TB/day maximum	• ~100 TB/day • Data may be transferred from ORNL, NERSC, ANL to NCAR, using Globus toolkit and successors.

10 Earth System Grid Federation: Federated and Integrated Data from Multiple Sources

10.1 Background

In climate science, the quantity of data in use by 2020 is expected to be in the hundreds of exabytes[1] (EB, where 1 exabyte = 10^{18} bytes). Current and future heterogeneous climate data will be distributed around the globe and must be harnessed to find solutions to mission-critical problems. Additionally, more requirements and more constraints are needed to expand and integrate new modeling capabilities and tasks, such as climate prediction, uncertainty quantification (UQ) of model performance, testbed development, and assimilation of more diverse data sets. These data exploration tasks can be complex and time-consuming, and frequently involve numerous resources spread throughout the modeling and observational climate communities. Staff expertise and core competencies must therefore be flexibly applied to multiple projects and programs to accommodate more complex applications and state-of-the-science analysis, while allowing resources to be adapted to address future areas of interest in climate research.

To process state-of-the-science models and analyze those results, researchers will need more complex and flexible architectures that can run heterogeneous applications over fast heterogeneous networks. Performing remote operations will reduce data movement and minimize the amount of data to be stored. By working closely with the community, the U.S. Department of Energy (DOE) Office of Biological and Environmental Research (BER) is exploring and developing hardware and software workflow applications to integrate DOE's climate modeling and measurements archives. BER is coordinating efforts to develop infrastructure for national and international model and data comparisons. Having an integrated infrastructure (or framework) in place will allow climate science centers worldwide to deploy a wide range of data visualization, diagnostic, and analysis tools with familiar interfaces — a critical issue for building data systems that process very large, high-resolution climate data sets and meet the growing demands of this intensely data-rich community.

Because the infrastructure for this type of community-wide network must be interoperable, each system must be established on a standard set of services, application programming interfaces (APIs), and protocols so that other systems can interconnect their components. By encapsulating component service operations behind message-oriented service interfaces, users will be isolated from the details of implementations and distributed service locations and freed to work in a virtual workspace, if so desired.

---

[1] CCDC Workshop, International Workshop on Climate Change Data Challenges, June 2011, http://www.wikiprogress.org/index.php/Event:International_Workshop_on_Climate_Change_Data_Challenges. CKD Workshop, Climate Knowledge Discovery Workshop, March 2011, DKRZ, Hamburg, Germany, https://redmine.dkrz.de/collaboration/projects/ckd-workshop/wiki/CKD_2011_Hamburg.

To this end, DOE has made elegant architectural investments in the Grid Forum: (1) Grid Computing and (2) Data Grids. In the area of Grid Computing, the Open Science Grid (OSG) successfully engages a variety of science domains in adapting their software components to use a distributed set of DOE and community-wide computing and storage resources. In the area of Data Grids, the Earth System Grid Federation (ESGF) (shown in Figure 12) successfully provides distributed data systems and services to the climate modeling community.

ESGF is an exemplary use case showing how to achieve an infrastructure with open, common-component architecture for distributed science data systems and services. For this infrastructure to have continual success, we must identify the overlapping needed

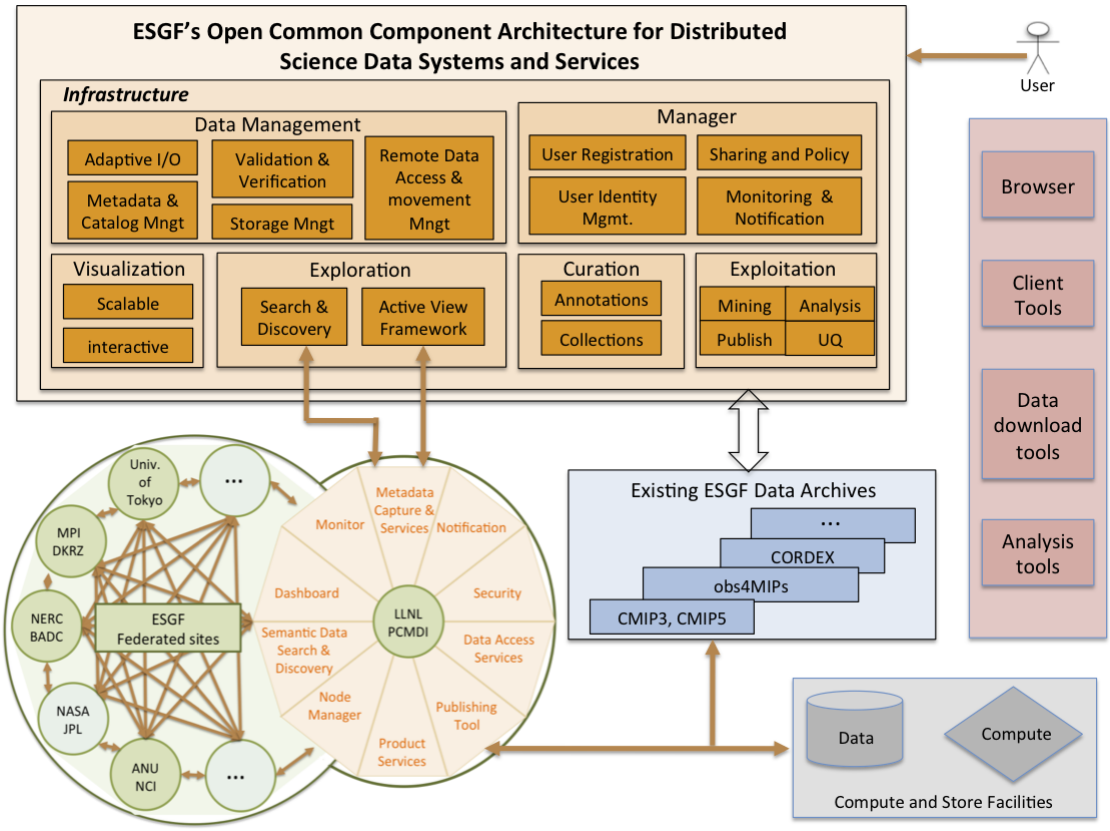


Figure 12. ESGF's federated framework integrates distributed data systems and services for discovery-class research that explores cross-cutting climate science domains. The dark orange boxes are common component services needed for the distributed data systems. Communications between the components are implemented via a standard set of APIs and protocols defined by the science community. ESGF currently comprises over two-dozen nodes, and five of these (indicated in the lower left by a darker shade of green) host replicas of a substantial number of the CMIP5 data sets (i.e., PCMDI, DKRZ, BADC, NCI, and the University of Tokyo). Users have access to all data throughout the federation, regardless of which ESGF node is used.

services that are common across project domains and design the system so that the integration works well and components are reusable, scalable, and extensible. As technology changes, the infrastructure must be flexible enough to evolve and keep pace with the growing demands of climate science.

Managing high volumes of climate data within a network of high-performance computing (HPC) environments presents unique challenges. How data are organized can have considerable impact on performance. Often, the only realistic choice for storage device is robotic tape drives within a hierarchical management system. In the worst case, poor data organization means that some data may never be accessed simply because it takes too long (i.e., storage access, compute resources, network). Users generally store the data themselves, possibly unaware of how to most effectively use the hierarchical storage system. Additionally, some applications require that large volumes of data be staged across low-bandwidth networks simply to access relatively small amounts of data. Finally, when data usage changes or storage devices are upgraded, large data sets may need to be reorganized and quite possibly moved to a new location to take advantage of the new configuration. To address these concerns and others, the climate data community needs a network architecture that offers more intelligent and complete layered data services (as shown in Figure 13), providing users with increased information about the data, its anticipated usage, storage requirements, and network system characteristics. Such a system will include a layered service structure that is invisible to users but that effectively manages the system to ensure a truly efficient, productive workflow.

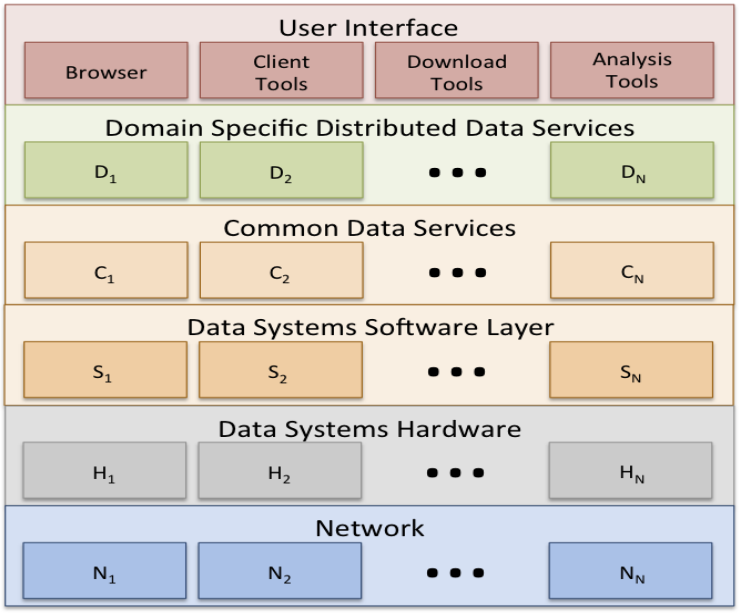


Figure 13. Diagram of the service layers hidden to the user. Standard APIs and protocols define the communication between each layer.

- **Domain-Specific Distributed Data Services.** At this level of the hierarchy, application components required for specific climate projects are clearly defined. For example, if a climate project has unique node protocol services for managing its distributed data system worldwide, those common components will fall into the first box of Figure 13. These services are specific to the domain space for this particular climate data project. D_1 to D_N captures the set of specific services for each project domain.
- **Common Data Services.** These are services that all project domain areas can use, such as moment, curation, discovery, annotation, and exploration. The C_1 to C_N layer exhibits standard protocols or standard interfaces from one layer to the next, allowing for extensibility and reuse.
- **Data Systems Software Layer.** The lower layers of the hidden services are closer to hardware and thus require definitions for more specific services. S_1 to S_N concerns itself with metadata, file size, provenance, and workflow.
- **Data Systems Hardware.** The H_1 to H_N layer represents hardware, such as clusters, clouds, and in situ data analysis for large-scale computational data analysis and modeling. At this level, interfaces must be defined to communicate with machines throughout the evolution of a simulation, for example, during the complete calculation for an uncertainty quantification analysis.
- **Networks.** Binding the collection of disparate hardware components, resources, and users are the networks. The N_1 to N_N represents high-speed (or low-speed) networks required to replicate large data holdings at storage facilities and to federate connectivity.

Given the critical importance of scientific climate data and its projected size by 2020, the climate research community must continue to specify a format for common activities as well as standards, APIs, and protocols to facilitate the development of infrastructures (such as ESGF) that support the community's long-range efforts. We cannot afford to work in an ad hoc fashion without proper standards for building hardware and networks or bonding software together via the specific protocols; doing so will cost DOE BER and the climate community at large considerable time and resources. If DOE is to optimize its investment in data, it must ensure that a common open architecture is in place. A significant fraction of that architecture is shared among the different climate activities, rather than having a specific domain architecture for each project.

10.2 Key Local Science Drivers

The Earth System Grid (ESG) was established in 1999 to meet the needs of modern-day climate data centers and climate researchers. Specifically, ESG addresses the requirements of data centers and climate researchers for interoperable discovery, distribution, and analysis of large and complex data sets. Under the leadership of the DOE BER Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory (LLNL) and in partnership with Argonne National Laboratory (ANL), the National Center for Atmospheric Research (NCAR),

Lawrence Berkeley National Laboratory (LBNL), Oak Ridge National Laboratory (ORNL), Los Alamos National Laboratory (LANL), the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and others in the national and international communities — including centers in the United Kingdom, Germany, France, Italy, Japan, and Australia — an internationally federated, distributed data archival and retrieval system was established under the name Earth System Grid Federation, or ESGF. Although this development effort is coordinated internationally, the ESG team is the primary contributor to the ESGF software stack. ESGF work has resulted in production of an ultrascale data system, empowering scientists to engage in new and exciting data exchanges that could ultimately lead to breakthrough climate-science discoveries.

ESG was critical to the successful archiving, delivery, and analysis of the Coupled Model Intercomparison Project (CMIP), Phase 3 (CMIP3), which provided data for the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC). It is proving to be equally important in meeting the data management needs of CMIP, Phase 5 (CMIP5), which is providing petascale data informing the 2013 IPCC's Fifth Assessment Report (AR5). Although ESGF has been indisputably important to CMIP, its current and future impact on climate is not limited only to this high-profile climate project. ESGF has been used to host data for a number of other climate projects, including the Community Climate System Model (CCSM), the Community Earth System Model (CESM), the North American Regional Climate Change Assessment Program (NARCCAP), the Carbon Land Model Intercomparison Project (C-LAMP), the Parallel Ocean Program (POP), and the Atmospheric Model Intercomparison Project (AMIP). These data archives have been augmented with observational data sets (for example, the Atmospheric Radiation Measurement Best Estimate [ARMBE], the Carbon Dioxide Information and Analysis Center [CDIAC], NASA satellite observation data sets [CloudSat, Microwave Limb Sounder (MLS), the Multi-angle Imaging SpectroRadiometer (MISR), the Atmospheric Infrared Sounder (AIRS), and the Tropical Rainfall Measuring Mission (TRMM)], and NASA–NOAA reanalysis data sets [Modern Era Retrospective Analysis for Research and Applications (MERRA), Clouds and the Earth's Radiant Energy System (CERES)]).

Because of rapid increases in technology, storage capacity, and networks and the need to share information, communities are providing access to federated open-source collaborative systems that everyone (including scientists, students, and policymakers) can use to explore, study, and manipulate large-scale data. The ESGF software stands out from these emerging collaborative knowledge systems in the climate community along multiple dimensions: the amount of data provided (petabytes), the number of global participating sites (dozens), the number of users (more than 25,000), the amount of data delivered to users (over 2 petabytes), and the sophistication of its software capabilities. ESGF is therefore considered the leader for both present and future data holdings as shown in the table below.

Table 7. ESGF distributed data archive

Type	Federated Data Sets (i.e., Projects)
Model	Phases 3 and 5 of the Coupled Model Intercomparison Project (CMIP3 and CMIP5)
Model	Coordinated Regional Climate Downscaling Experiment (CORDEX)
Model/Observational	Climate Science for a Sustainable Energy Future (CSSEF)
Model	European Union Cloud Intercomparison, Process Study & Evaluation Project (EUCLIPSE)
Model	Geo-engineering Model Intercomparison Project (GeoMIP)
Model	Land-Use and Climate, Identification of Robust Impacts (LUCID)
Model	Paleoclimate Model Intercomparison Project (PMIP)
Model	Transpose-Atmospheric Model Intercomparison Project (TAMIP)
Observational	Clouds and Cryosphere (cloud-cryo)
Observational	Observational Products More Accessible for Coupled Model Intercomparison (obs4MIPs)
Model	Reanalysis for the Coupled Model Intercomparison (ANA4MIPs)
Model	Dynamical Core Model Intercomparison Project (DCMIP)
Model	Community Climate System Model (CCSM) / Community Earth System Model (CESM)
Model	Parallel Ocean Program (POP)
Model	North American Regional Climate Change Assessment Program (NARCCAP)
Model	Carbon Land Model Intercomparison Project (C-LAMP)
Observational	Atmospheric Infrared Sounder (AIS)
Observational	Microwave Limb Sounder (MLS)

10.2.1 Instruments and Facilities

LLNL computational hardware and networks are supported by Livermore Computing (LC), which delivers a balanced HPC environment with constantly evolving hardware resources and a wealth of HPC expertise in porting, running, and tuning high-bandwidth, large-scale applications. Currently, LC delivers multiple petaflops of compute power, massive shared parallel file systems, powerful data analysis platforms, and archival storage systems that can hold many petabytes of data. This balanced hardware environment supports key collaborations between LLNL applications developers and LC experts on the creation, debugging, production use, and performance monitoring of HPC parallel applications, as well as data analysis in a variety of scientific disciplines, including climate science. PCMDI's collaborations with LC represent LLNL's first grid computing project involving international collaboration.

Figure 14 shows all PCMDI/ESGF components within the boundary of the so-called Science Demilitarized Zone (DMZ). Two major storage areas hold petabytes of climate data: (1) the Green Data Oasis (GDO) and (2) the newly acquired Climate Storage System (CSS). GDO and CSS are part of the Green Collaboration Environment (GCE) network, which is managed by LC and has its security covered by the GCE network core service description. GCE was established specifically for PCMDI's international collaboration needs.

In terms of ensuring good performance with GDO and CSS via Grid services (such as GridFTP), a set of Linux servers (indicated by 1 in Figure 14) has back-end connections

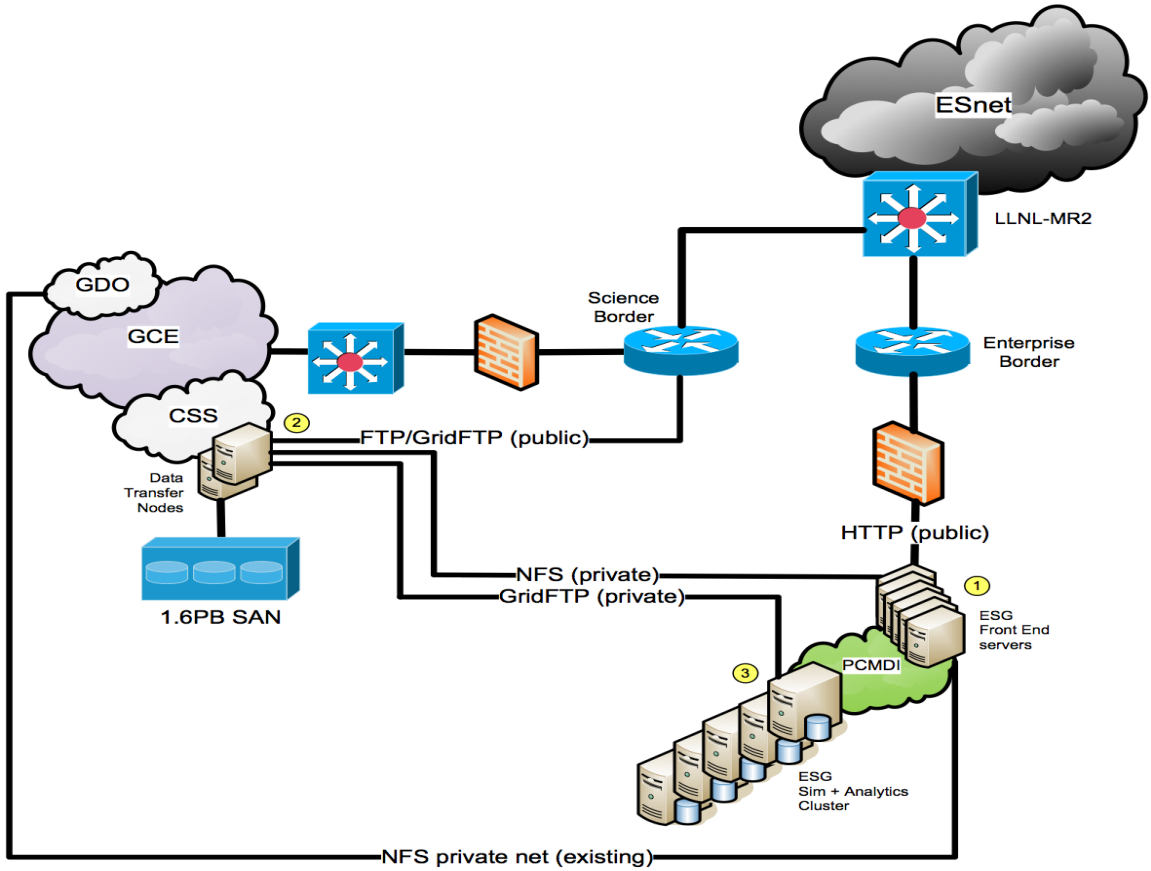


Figure 14. Data flow for ESGF portals: (1) Users communicate with ESGF front-end servers on the LLNL Green Network via HTTP. Small data sets are retrieved by the front-end nodes from the CSS SAN over a private NFS network and returned to the user via HTTP. (2) Large data sets are made available to users directly from the CSS storage system's data transfer nodes (DTNs) via GridFTP. Firewall bypass on the DTNs is required to ensure good/consistent performance for these large file transfers. (3) ESGF may perform analysis of raw data if requested by users through the front-end servers. Analysis jobs are dispatched to the ESGF Sim+Analytics cluster, which retrieves necessary raw data from CSS SAN over a private network via GridFTP, performs analysis, and puts results back in CSS SAN, where they can be retrieved by the customer via GridFTP.

via the network file system (NFS) to the GDO and CSS Solaris servers. In addition, the back-end ESGF clusters (indicted by 3 in the diagram) are Linux servers connected to the storage devices via private GridFTP connections.

10.2.2 Process of Science

As the demand for robust and consistent scientific data distribution platforms increases, user interface (UI) design and implementation have become one of the more important tasks in the creation of a scientific data access portal. Potential data consumers, specifically end users of the ESGF portal, are inevitably concerned with the manner in which services (such as node management, search, and login) are presented. For the most part, end users discover and access data via the ESGF front-end UI application. The ESGF node team has incorporated a number of components to the front-end design to

Figure 15. The ESGF node installation script deploys the home page of a minimally configured Web front-end application, shown at left. Included on the home page are search capabilities, information pertaining to the node and organization, quick links to commonly used tools, links to other ESGF nodes, and analysis tools. Example search page and analysis/visualization results are shown on the right.

accommodate a wider community of potential users who may be interested in various data sets (e.g., scientists interested in comparing modeled output with observational data). As with most modern search portals, text search has been enhanced with

autocomplete capabilities to aid users who may be unsure of specific search strings. The temporal search tool leverages the highly effective range search capabilities of the Apache Solr search back end, allowing users to extract data sets according to ranges constrained by time measurements via Solr's Lucerne-based inverted-index technology. The faceted-based category navigation tools were expanded to include flexible terminology (for example, "instruments" in observational terminology) while providing direct support for the structured terminology of the climate community.

As shown in Figure 15, UI components have been developed for the key functional areas listed below. Significant refinement has been achieved in each component area based on production use of the system and user feedback. The design goals for UI tools include exposing the base system functionality, providing a consistent and intuitive user experience, and supporting a flexible and maintainable framework for future enhancements and revisions.

- **The home page** provides visitors with general information about the discipline-specific portal, starting points for discovering data collections, direct access to notable data collections, important notices regarding system status, and access to login and account-request functions. The home page allows a climate project to customize its format to include project-specific information, data browser entry points, logo images, and color palette.
- **User registration** offers a multistep workflow for account creation, approval, and validation. The resulting account may be used to authenticate any user at any portal in the federation.
- **User and group management** allows registered users to change account settings and request access to privileged data collections. It also provides tools for group administrators to approve group requests and manage group membership.
- **Login** allows registered users to authenticate with the federated system with an OpenID user identifier. Users may request password delivery via e-mail in case of lost credentials.
- **Data browsing** provides support for file system-like hierarchies and high-level associative arrangements such as experiment- and project-related listings.
- **Data search** is the primary data-discovery method for most users. This component provides a simple and familiar text-based search as well as faceted navigation for exploration-based metadata inquiry.
- **Data download** allows for individual file download via hyperlink and for bulk download requests from the file-listing interface using generated wget scripts and the Globus Online-hosted data movement service. If data collections are restricted and under access control, the user is directed to authenticate prior to data download.
- **Data transfer** allows registered and authorized users to request and manage groups of files from deep storage systems throughout the federation. Users can access real-time status reports and are notified by RSS feeds or via e-mail when transfers are complete.

- **Data visualization and subsetting** provides an interface for requesting charts, plots, and data subset downloads. Users may choose variables of interest and select subregions geospatially using an interactive map and temporally using time controls.

Users can also interact with the ESGF distributed archives for knowledge discovery via analysis tools. Funded under BER in support of its science mission, the Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) framework is designed to integrate six analytical and visualization tools — CDAT, VisTrails, ParaView, VisIt, DV3D, and R — all under one application. Based on Python, it links disparate software subsystems and packages to form an integrated environment for analysis. UV-CDAT's design and openness permit the shared development of climate-related software by the collaborative community. In particular, the goals of the UV-CDAT project are to (1) prepare for the CMIP5 data archive and assessment process by developing derived data products and user-reproducible workflows and analysis archives; (2) develop capabilities to inter-compare ungridded observational data sets and model data for validation;

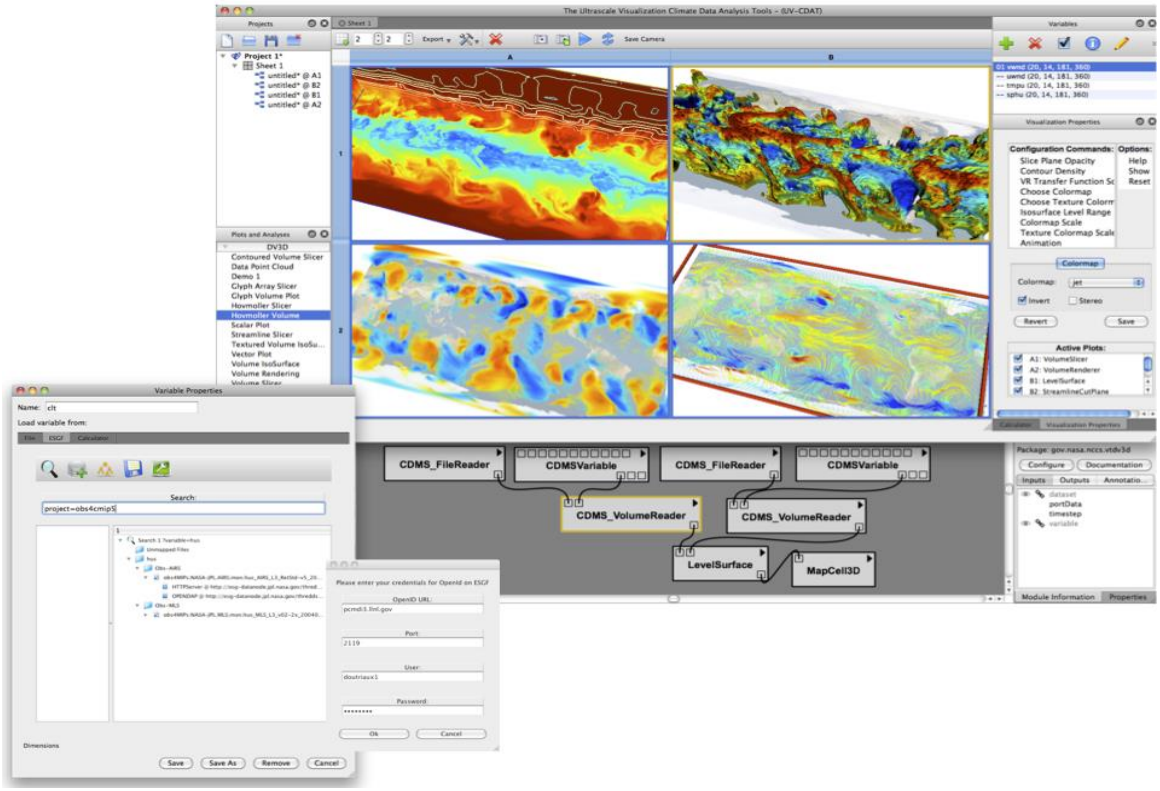


Figure 16. UV-CDAT searches and accesses the ESGF node archive at LLNL/PCMDI (shown in the lower left) and displays the requested results using DV3D (in the upper right). The UV-CDAT reproducible workflow is displayed under the four-panel visualization spreadsheet.

(3) deliver efficient scalable analyses and visualization for high-resolution simulation data; (4) deliver data products in formats suitable for expert and non-expert users; and

(5) build all capabilities on existing ESGF node infrastructures. Figure 16 shows the new interactive UV-CDAT graphical user interface (GUI) in the background and the UV-CDAT ESGF node search and browse GUI. These features allow users to search and browse the ESGF distributed archive from within the UV-CDAT analysis tool as if they were on a Web browser. Once a data set is located, a user can download it directly to the UV-CDAT application.

The interaction between users and the UV-CDAT GUI is also depicted in Figure 16. A user interacts with the UV-CDAT GUI by invoking scripts, clicking buttons, or dragging variables and plot types. In response to these actions, UV-CDAT records a series of operations and converts them into provenance-enabled workflow operations that allow the user to share work with others as well as to reproduce the operations.

10.3 Key Remote Science Drivers

ESGF seamlessly joins climate science data archives and users around the world. As shown in Figure 18, it accesses many wide-area networks (WANs) to remotely connect researchers, policymakers, and other users to climate data projects through Web-based interfaces and analysis tools (as described in Section 10.2.2). Data providers make data available to the federation by publishing to one of two-dozen ESGF node portals. Data can be replicated at other ESGF node sites for backup, to improve ease of use, or to exploit site resources. In the process of ESGF node replication, data are moved via GridFTP or via the HTTP protocols. This process is the same for user data movement.

Part of the process of publishing and replicating data is the data quality control (QC) check operations, which ends in digital object identifiers (DOIs). This process takes a three-layer approach to data-quality assurance, as shown in Figure 17. When a modeling or data center publishes data to ESGF, the system performs the automatic QC check level 1 on the data. This QC check confirms that the data are in compliance with the netCDF and Climate Forecast (CF) convention. For the second-level QC check, subsets of data are transferred to either DKRZ (MPI) or LLNL, where a QC code is run on the data to check for consistent application of units, measurements, etc. In some cases, visual inspection of the data also takes place for correctness. Once the data are accessible to the user community and have been used without complaints for a three-month period, they are elevated to QC level 3 status and issued a DOI. DOI data sets are then replicated to the IPCC Data Distribution Centre (DDC) at the World Data Center for Climate (WDCC) and to long-term archive at DKRZ and BADC. This process will occur over the next two years.

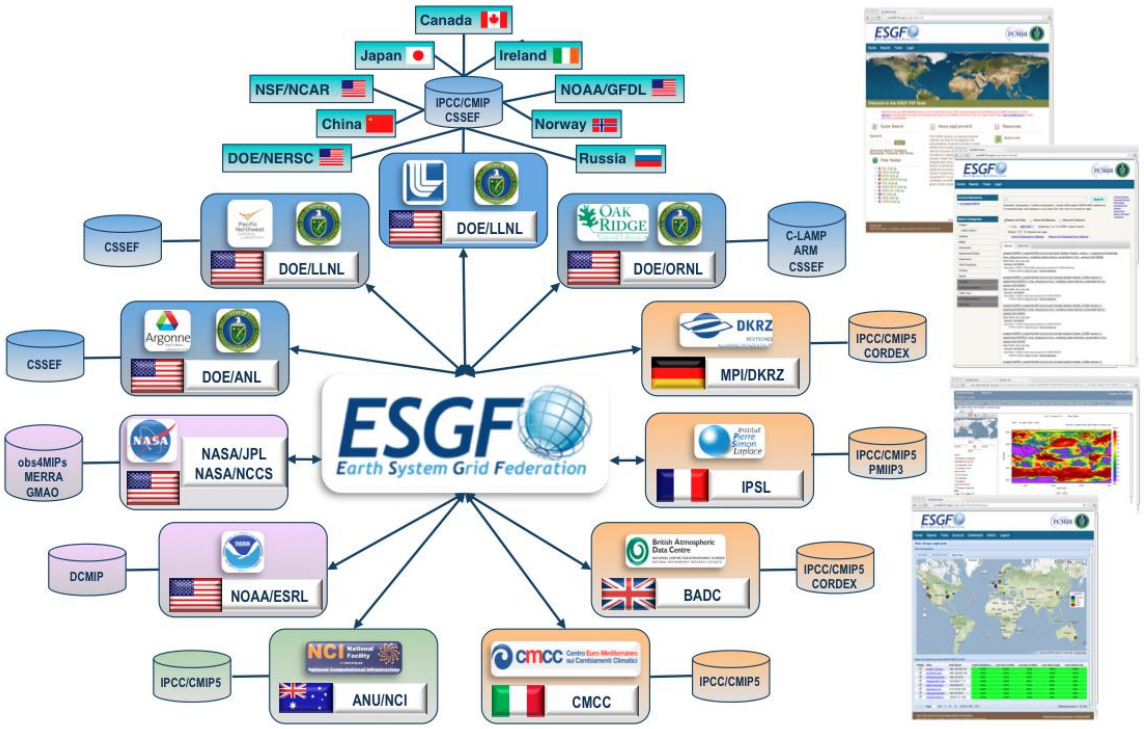


Figure 18. Federation of ESGF showing collaborations between a few of its remote data centers, data archives, and potential data transfers between sites. For example, the U.S. DOE/LLNL portal (at the top) harvests IPCC/CMIP5 data from 10 countries. That is, the original data resides at the data centers, but subsets of the data are replicated at LLNL for backup, greater access, and use. The DOE/LLNL portal URL is http://pcmdi9.llnl.gov.

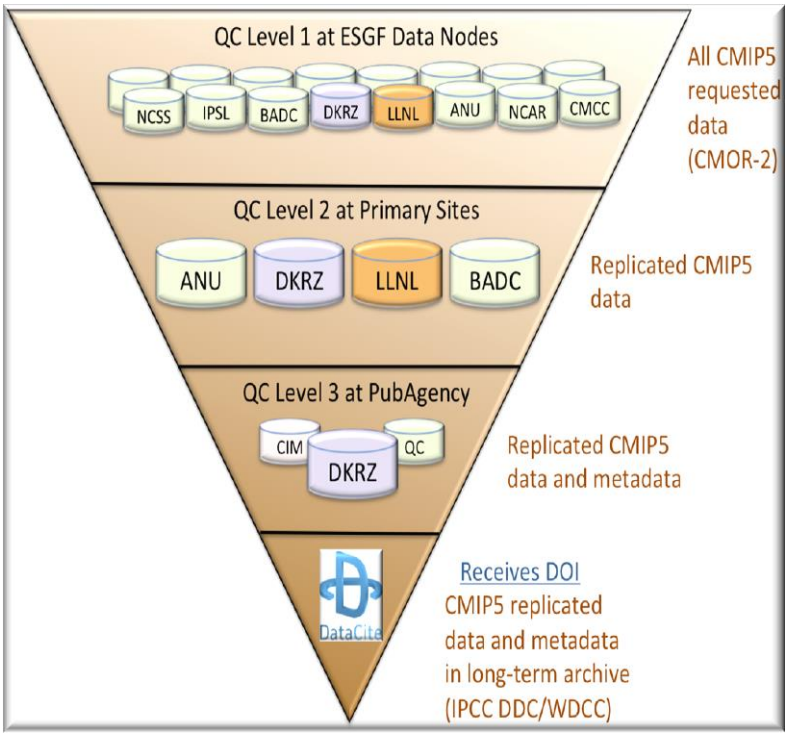


Figure 17. Three-layer quality assurance concept

10.4 Local Science Drivers for the Next 2–5 Years

Rapid advances in experimental capabilities, networks, hardware, computational technologies, and techniques are driving exponential growth in the volume, acquisition rate, variety, and complexity of scientific data. This new wealth of meaningful data has tremendous potential for scientific discovery. However, if scientists are to use this vast resource to achieve scientific breakthroughs, the data holdings must be exploitable so that the information can be analyzed effectively and efficiently, and the results shared and communicated easily. The explosion in data complexity and scale makes these tasks exceedingly difficult to achieve — particularly given that an increasing number of climate projects are working across techniques, integrating simulations with experimental or observational results.

Consequently, we must continually build on ESGF's data-management, analysis, and visualization tools to provide research teams with easy-to-use, end-to-end solutions. These solutions must facilitate (and where feasible, automate) every stage in the data life cycle, from collection to management, annotation, sharing, discovery, analysis, and visualization. Hereby, a core set of ESGF functionalities must be offered to all climate projects, but individual processes will require customization so they can be adapted to a project's specific needs and fit into the different research and analysis workflows.

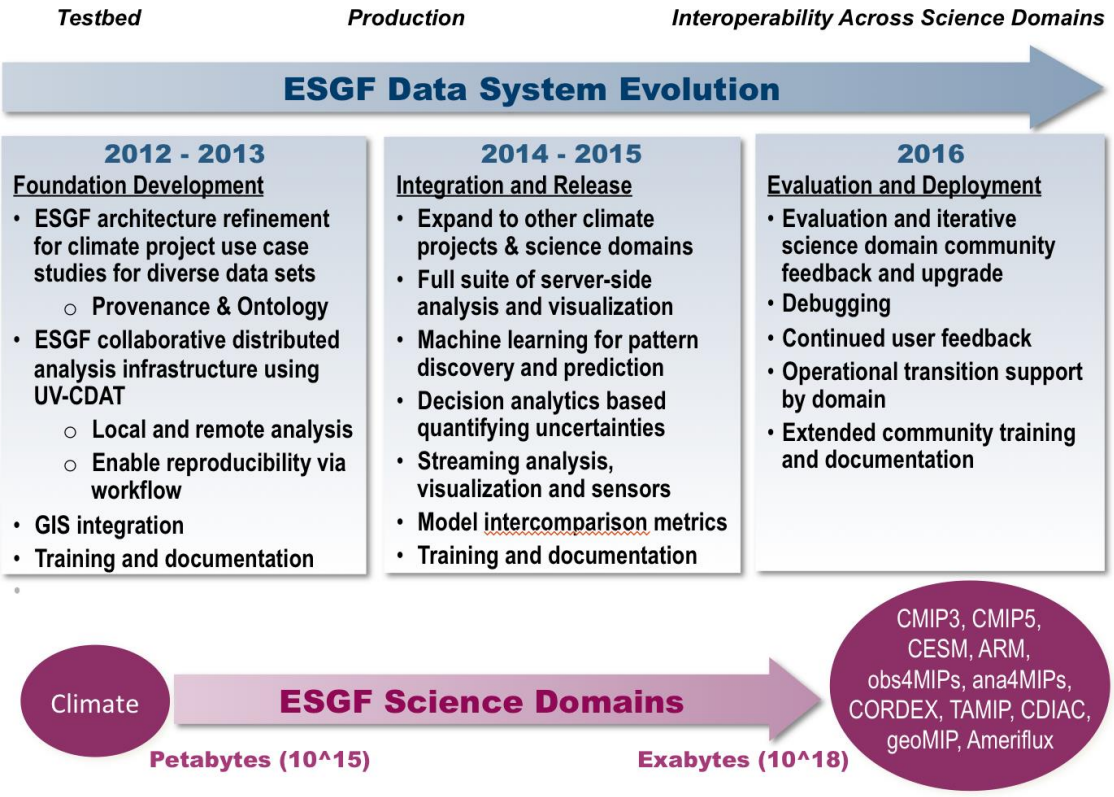


Figure 19. High-level road map for evolving ESGF across many climate-science projects

Therefore, we will leverage existing DOE and community-proven core technologies and facilities so that we can provide an even more comprehensive portfolio of data-management, analysis, and visualization capabilities to the entire climate community. We will build on technologies developed within DOE-funded projects — such as the ESGF, UV-CDAT, Globus Online, CSSEF test bed, Ensemble Data Analysis Environment, Sapphire, and Extreme Scale Visual Analytics — and adapt and extend these tools in collaboration with DOE application partnerships (U.S. and international agencies, universities, and private companies).

In addition to our trademark advances in data preparation, search and discovery, data access, security, and federation (as mentioned above), over the next few years we will focus on expanding into the new areas of data mining, provenance and metadata, workflows, HPC, data movement, and data ontology. Using projections of upcoming scientific endeavors, we can extract and summarize the high-level requirements that we plan to address for ESGF, as shown in Figure 19.

10.4.1 Instruments and Facilities

The increasing data volume generated by climate science, coupled with the new capability provided by the DOE ESnet 100 Gbps initiative, requires a commensurate solution to manage and deliver extreme-scale data sets over advanced networks and not-so-advanced networks. Today's storage and data transfer solutions perform poorly when file sizes vary significantly, which is common in the large data sets produced by climate simulations and observational research. In the near future, climate scientists will confront an exacerbating situation in the emerging network because large networks require larger data sets and incur even greater variance in file sizes. Furthermore, the file system, storage facility, and switch fabrics as a whole cannot provide an input/output (I/O) throughput commensurate with the emerging network speed. Developing an efficient transport coordination protocol for substandard to advanced networks to manage disks, storage units, file systems, and network buffers remains an unexplored territory because of its complexity. Over the next two years, we will need to examine the coming 100 Gbps network technologies coupled with ESGF's extreme-scale data-management tools to deliver data ranging in size from many terabytes (10^{12}) to tens of petabytes (10^{15}). This effort will incur additional hardware storage costs along with HPC and cluster computing costs.

10.4.2 Process of Science

In the coming years, the process of conducting data-intensive climate-science research will remain primarily the same as described in Section 10.2.2; however, data processing will be more commonly performed at remote data centers. The goal is to have UV-CDAT analysis processes co-located where the data reside. Through UV-CDAT, provenance metadata would be recorded at every step in the process and archived as a workflow configuration co-located with the data product. Later, other scientists could run the

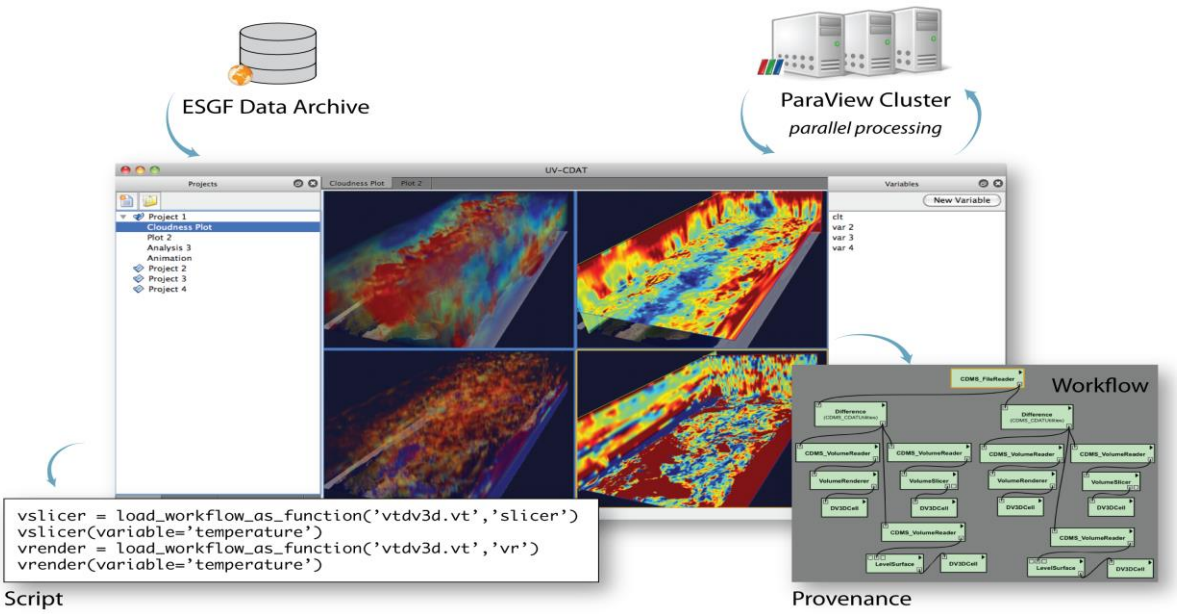


Figure 20. A glimpse of the ESGF infrastructure accessing distributed managed data at Leadership Computing Facilities. Parallel processing, performing data reduction and analysis, takes place via ParaView cluster analysis, and multiple views are displayed to the user. The workflow captures the entire process for reproducibility and knowledge sharing

same analysis from the workflow descriptor to confirm the results, and they could expand on early findings by running different variations of a processing algorithm or using different input data sources.

Figure 20 illustrates one view of the planned infrastructure for ESGF. In this future setup, key data centers, identified as Leadership Computing Facilities, are located throughout the federation, providing users with ready access to the complete data archive. A ParaView cluster performs the calculations requested for different projects and offers users multiple display options for viewing the returned results. All data products and workflow descriptors in this planned ESGF infrastructure would be automatically archived to improve the ease of sharing knowledge, both about the climate predictions and the data-analysis applications.

10.5 Remote Science Drivers for the Next 2–5 Years

Because ESGF is an international collaboration of partners, the remote science drivers are the same as the local science drivers described in Section 10.4. That is, partnering activities will involve many U.S. and international agencies, universities, and private companies.

10.5.1 Instruments and Facilities

ESGF data sources are distributed among national and international data repositories (as shown in Figure 18). In the process of scientific data analysis, project teams frequently access a large portion of the data archive and move a large volume of data from one repository to another. Our current plan is to analyze the data at the local computing facility where it is stored. However, this approach does not scale to the extremely large data sets encountered in the extreme-scale sciences. For example, in climate research, the comparison between observational data, which is recorded at a very high sampling frequency, and Earth simulation data with high resolution, which comes from different data repositories than observational data repositories, requires a vast amount of data to be moved to a large computational facility — such as LLNL. One challenging issue in such an effort is the limited ability of large networks to manage, manipulate, and explore data generated at ultrahigh space and time resolution. Although DOE is investing in large-scale networks (i.e., ESnet's 100 Gbps network), the best process for using these new resources is still to be determined. Furthermore, the

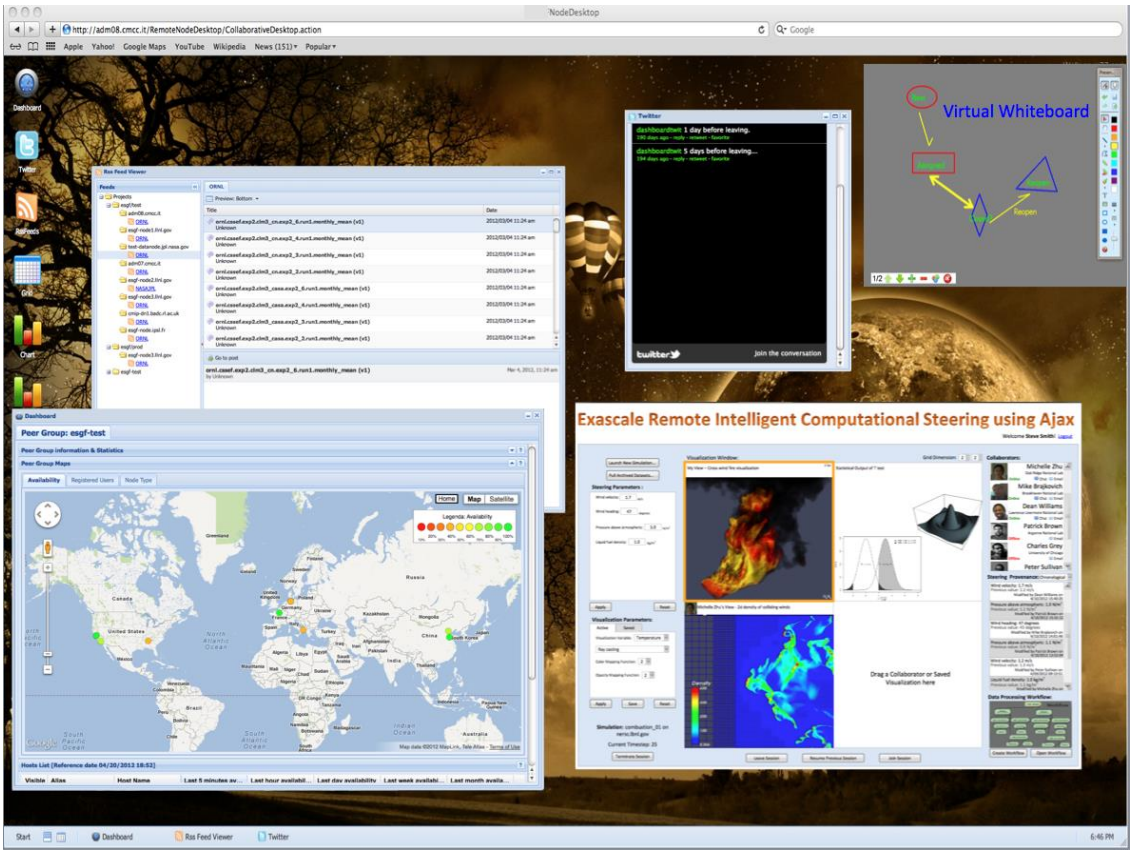


Figure 21. A mockup shows the projected ESGF online environment as tools and capabilities are expanded. Users would have access to search results, dashboard, and other collaborative desktop tools (such as Twitter and RSS feeds and a whiteboard-type working environment for collaborative interactions).

models of data flow over high-capacity networks for these extreme-scale sciences are expected to be quite different from the traditional workflow approaches deployed in less data-intensive sciences, which are designed to overcome current network limitations. These limitations and performance uncertainties have led to a significant underutilization of the available high-bandwidth resources.

10.5.2 Process of Science

Given the current state, challenges, and demands on users, we believe the most productive approach to serving a community of collaborative users is to develop a Web-based system that offers remote visualization and online steering capabilities (in particular, via mobile application-handled devices) in the next two to five years. Furthermore, such a system should support dynamic and intelligent scheduling and mapping to minimize the complete end-to-end delay or maximize the frame rate that users may experience because of the data volume and complexity required for exascale sciences. For example, when navigating across climate projects, a user usually encounters new sets of difficulties: Each project has its own set of tools, and each tool operates through customized features such as language, data structures, and hardware requirements. These wide-range toolkits render collaboration across climate projects nearly impossible. The simple task of accessing, let alone actually using, data can be so challenging that it will crush a multi-institutional effort in its infancy. Having an integrated Web-based, component-based system will alleviate most of the difficulties and will enable new partnerships that are not even conceivable at this time.

Because ESGF is a distributed system, the reporting process used in climate science research over the next 2 to 5 years will continue to be similar to the process described in Section 10.4.2.

10.6 Beyond 5 Years — Future Needs and Scientific Direction

It is difficult to project what resources will be required in the next 5+ years for the continued development of ESGF to meet the changing needs of the climate science community. We are, however, working toward building and delivering an implementation that can transfer petabytes of data to designated facilities as a routine event. New capabilities for extreme-scale data movement will allow researchers to more effectively use the full resources available with the DOE ESnet 100 Gbps network infrastructure coupled with upgrades in computing hardware and integrated software such as ESGF. More specifically, we are moving toward:

1. Data-set-level streaming protocol using the ESnet 100 Gbps network
2. Quantitative analysis on how best to apply the 100 Gbps network for ultralarge sets of climate data as well as other extreme-scale applications on data
3. Evaluation of the performance gap between traditional and faster advanced networks and the underlying storage systems
4. Continued vetting of technology for climate research such as ESGF, in particular, to ensure that the petascale-exascale data sets are constantly synchronized

among the four international climate data centers, the DOE laboratories, and other U.S. agencies

10.7 Network and Data Architecture

Section 10.2.1 describes the components planned for DMZ and the direction of LLNL's data architecture. Figure 14 shows the data architecture in design today, which will provide the petabytes of disk storage needed for various data products, such as CMIP5 and statistical downscaling data. Trends for data-intensive computing, especially in extreme-scale communities such as climate science, indicate that computational needs will continue to grow into the future, in much the same as they have up to today, but on a much larger scale at all levels of network architecture to deliver and store the coming exascales of data products.

10.8 Collaboration Tools

Our focus on partnerships and collaborations has led to close relationships with a wide variety of data, science, and technology efforts. These relationships have positioned ESGF to make important contributions to the progress of science in CMIP5, CSSEF, CESM, and other data-intensive climate-relative community projects, as mentioned in Section 10.2.

To effectively build a distributed infrastructure that can accommodate the needed petascale-exascale data management and analysis enterprise, the ESGF team established connections with researchers and scientists involved in other national and international climate programs aimed at assisting in DOE's climate mission. These involved discussions — through workshops, conferences, and face-to-face meetings — with researchers at many other centers and institutes and throughout the climate community, all of whom have a strong interest in collaborating on ESGF projects. In most cases, these collaborative efforts held weekly team meetings (via e-mail, WebEx, GoToMeeting, Skype, or teleconferencing) to discuss progress and technical issues.

10.9 Data, Workflow, Middleware Tools, and Services

ESGF delivers a comprehensive, end-to-end and top-to-bottom environment for current and emerging exascale climate science, as shown in Figure 22. We emphasize data services at each level of the architecture. Figure 22 is an expanded view of Figure 13, showing in greater detail the hidden-layer services along with the analysis and visualization services accessible to users. More than a proof-of-concept, the production of ESGF for climate projects is evidence that a distributed dynamic federated system is flexible enough to support a wide range of climate projects by providing the following current and futuristic capabilities:

1. Federated heterogeneous data architecture framework
2. Service-oriented and layered architecture
3. Application layers, offering domain-specific services and data portability

4. Common services layer, such as data access, discovery, replica selection, task management, virtual data catalog, remote computation, remote visualization, and remote sensors
5. Data systems software layer, with such information as metadata, formats, semantic standards, ontology, replica catalog, and security protocols
6. Data systems hardware, including storage systems, clusters, Leadership Computing Facilities, and display devices
7. Networks and the related services, including virtual networks, network caches

10.10 Outstanding Issues

As our work on ESGF shows, building the infrastructure for extreme-scale computing and gathering support from the research community to sustain a distributed network are significant challenges. To continue to build on the successes of ESGF, we recommend that DOE BER host a data forum, where data systems and services architects from each of the national and international climate projects can discuss methodologies, philosophies, and standards common for all. The goal of such a forum would be to establish an open, common-component architecture for distributed science data systems and services within the greater community. However, the forum should not be limited to DOE, but open to the entire federation (including U.S. agencies such as

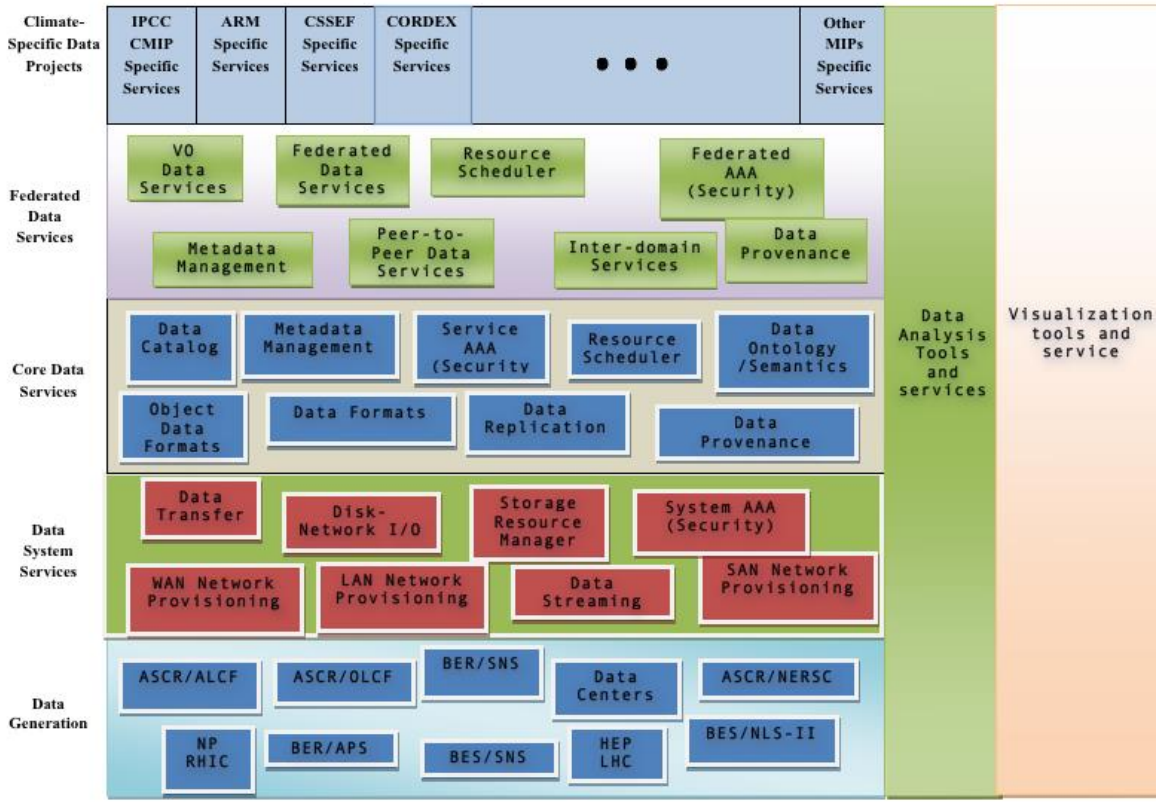


Figure 22. Current and future end-to-end infrastructure for ESGF shows the framework and relationships for distributing climate science data and services

the National Science Foundation, NASA, and NOAA and the international partners) to encourage the large-scale adoption of all approved standards — a long-term goal that DOE cannot accomplish in isolation. For this important national endeavor, which involves distributed data organization, archiving, and sharing of dispersed resources, we encourage partnering by all organizations.

10.11 Summary

ESGF is a key data-dissemination infrastructure and resource for climate simulation, observation, and reanalysis data. We have three major activities that affect the need for increased bandwidth: receiving data from all of the providers (including periodic updates), replicating that data to other national and international sites, and responding to requests from users for portions of the data holdings. Each activity requires network bandwidth the size of a PB data repository. For the future ESGF archives, the repository will increase by many orders of magnitude, but at best, network bandwidth will increase by only 1 order of magnitude in the next 5+ years. By replicating data sites, we can spread out the demand for services, which will resolve part of the gap. However, all ESGF sites need the fastest available network as soon as possible if the federation is to succeed at delivering results to prospective customers and large-scale data movement.

The Summary Table in Section 10.12 describes the data moment for only one data project — CMIP5. It also describes only one process — moving data between sites for replication, mainly at LLNL by ESGF administrators. If this example were to include end-user download and the total federation of other projects, the transfer amount would quickly multiply by 2 or 10, depending on community activities.

10.12 Summary Table

Current science drivers and projected network needs for CMIP5.

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• ESGF distributed nodes worldwide: ○ Today, 25 ESGF node sites ○ In two years, 50 to 100 ESGF node sites	• 20 climate data-intensive projects: CMIP3, CMIP5, CESM, CSSEF, ARM, obs4MIPS, ana4MIPS, CORDEX, EUCLIPSE, GeoMIP, LUCID, PMIP, TAMIP, DCMIP, POP, NARCCAP, C-LAMP	• CMIP5 data only • Data volume ○ Today, 1.8 PB ○ In two years, 3.5 PB • Data-set composition ○ Today, 3.2 million files ○ In two years, over 6 million files	• Replicating CMIP5 data to LLNL only ○ 5 TB per day	• Replicating CMIP5 data internationally ○ 50 TB per day ○ Data are transferred to sites, LLNL, BADC, DKRZ, and U. of Tokyo
2–5 years				
• ESGF distributed nodes worldwide: ○ 100s of ESGF node sites	• More than 20 climate data-intensive projects, (estimate could reach as high as 100 projects or more)	• Total federated data volume ○ 100s of PB ○ 100s of millions of files	• Routinely replicating data ○ 10 PB/day	• International federation ○ 10s of PB per day ○ Data are transferred to multiple sites
5+ years				
• ESGF distributed nodes worldwide: ○ 100s to 1,000s of ESGF node sites	• More than 100 data-intensive projects (including other science domains such as biology, cosmology, chemistry, and materials science)	• Total federated data volume ○ 1 EB or more ○ 1,000s of millions of files	• Routinely replicating data ○ 100s PB/day	• International federation ○ 100s of PB/day ○ Data are transferred to multiple sites

Acknowledgments. This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

11 Easy, Reliable, Secure, High-Performance File Movement

11.1 Background

This case study encompasses a range of Biological and Environmental Research (BER)-relevant application scenarios, all involving the *movement* and/or *sharing* of large quantities of file data. We identify common requirements for networks, middleware, and tools; describe lessons learned meeting those requirements using the Globus Online software-as-a-service (SaaS) system; and identify gaps that should be addressed to improve treatment of these scenarios in the future.

11.2 Key Local Science Drivers

For simplicity, we group discussion of local and remote science drivers in the next section.

11.3 Key Remote Science Drivers

11.3.1 Instruments and Facilities

Other case studies provide details concerning major BER-relevant instruments and facilities. In brief, these include supercomputers (e.g., Argonne Leadership Computing Facility [ALCF], National Energy Research Scientific Computing Center [NERSC], Oak Ridge Leadership Computing Facility [OLCF]); genome sequencing facilities; and Basic Energy Sciences (BES) facilities such as light sources that are used for BER-related projects. These instruments and facilities vary widely in the capabilities that they offer to users. For example, Advanced Scientific Computing Research (ASCR) supercomputer centers provide substantial storage to their users (at least for the duration of a project) while BES facilities may require that data be removed from the facility at the end of an experiment.

Community data repositories have also emerged as an important player in both biological and environmental research. As the volume of experimental, observational, and simulation data grows, and the importance of that data for science increases, it becomes increasingly important to collate relevant data into curated data repositories: e.g., see Earth System Grid[1] and the Systems Biology Knowledgebase (KBase). Such data repositories may be centralized or, increasingly, distributed (e.g., Earth System Grid Federation [ESGF]).

Science projects often also make use of non-DOE facilities (e.g., NSF supercomputers, other sequencing centers, NASA data systems) and of local computing and storage facilities, e.g., at universities (often substantial in size). The use of cloud facilities by university collaborators is growing.

As we describe in the next subsection, many instruments and facilities are capable of producing and/or consuming large quantities of data, and large-scale data movement

among them is frequently important to the scientific process. The biggest obstacle to their effective use appears in many cases to be not the wide area network but the "last mile" — the various elements at the network endpoints that are required to enable reliable, high-speed, end-to-end transfer. See below for more discussion on this point.

11.3.2 Process of Science

We describe a set of BER-relevant data movement use cases that illustrate the range of file movement activities that can occur during scientific research. Such activities may involve purely local, a mix of local and remote, or only remote resources, depending on the application.

We provide some rough data size estimates in the following. File transfers range in size from megabytes to petabytes (the largest transfer performed with Globus Online to date, albeit not for a BER application, was by the Laser Interferometer Gravitational Wave Observatory [LIGO] project, which mirrored around 1 PB for fault tolerance) and the number of files from 1 to millions. Even "small" transfers (in terms of number of bytes) can be challenging to perform if they involve large numbers of files. As noted in the recent Terabits networking workshop[2] "the average file size on the OLCF parallel file systems is only 14.8 megabytes, indicating that the median file size is much smaller"; in a sample IPCC CMIP-3 data set, 70% of files are <200 MB, and 30% of the data files are <20 MB; and the new Dark Energy Survey expects its median file size to be only about 150 KB. Though research data are typically structured, and while technologies such as Hierarchical Data Format (HDF) are seeing broader use, for better or worse the file system is still frequently used as an organizational framework for data. Thus, we expect that small files will continue to be a problem into the future, and indeed an increasingly bigger problem as networks get faster. Studies in 1984[3] and 2006[4] of file size distribution in the same university computer science department showed that over 22 years, the median file size only doubled, from 1080 to 2475 bytes. Admittedly, the files in a DOE research project are likely to have different characteristics, but these data are suggestive of the problem.

Mirroring and archiving. A common reason for moving large quantities of data across both local and wide area networks is to create copies on other storage systems, including archival systems. This mirroring may be performed to improve access performance (e.g., as in the ESGF) or for fault tolerance and preservation. Within the DOE system, a common destination for transfers for archival purposes appears to be NERSC. Data sizes here can become large: tens or hundreds of terabytes and millions of files. Transfers to tape-based archival storage involve unique performance optimization challenges.

Movement due to storage system unavailability. Another surprisingly common reason for data movement is changes in storage system availability. For example, a storage system becomes full, a storage allocation expires, an archival system shuts down — all such occurrences can spur large-scale data movement to other facilities. The characteristics of such transfers are similar to those for mirroring and archiving.

Publication to data repository. As community data repositories grow in importance, so too do the activities relating to publication into these repositories. These activities can include data movement, quality-control operations, metadata extraction or synthesis, approval processes, and integration into catalogs. Individual publication operations tend to be smaller than in previous cases, involving GB or TB rather than tens or hundreds of TB, but can be large in aggregate — and rapid response time is needed. An emerging need here is for tools that handle file transfer, but also to automate the other dimensions of publication.

Download from data repository. Users inevitably want to download data from community data repositories for local analysis. As data repositories get larger, it becomes less feasible for users to mirror a data repository in its entirety (as users have often done in the past), but large-scale downloads clearly continue to be viewed as desirable.

Download from scientific facility. Having produced data at a scientific facility, researchers frequently want to transfer that data to another location (e.g., university computer facility, cloud provider) for storage and/or analysis. The frequency and importance of such transfers seems to vary a great deal among facilities: e.g., ASCR LCFs tend to provide a complete environment that allows data analysis to be performed locally (but that can still require intrafacility data movement), while other facilities require users to take data home. In genomics, we see (at least in universities) a growing number of researchers transferring data to cloud service providers (in particular, Amazon) for storage and analysis.

Data analysis. As data becomes larger and thus amenable to visual inspection, and researchers become more sophisticated in their analysis methods, it becomes more important to be able to apply computations to entire data sets rather than just to individual elements. To support such use cases, it must be possible to transfer data (or locate data permanently) near to computers. Thus, we find a frequent driver of file transfers is to move data to/from a local or remote computer system.

File sharing. An increasingly common user requirement is to share data with collaborators, project members, and/or the community at large. Sharing cannot require that the people with whom data are to be shared have an account on a DOE computer, as this introduces an inordinate barrier to sharing.

11.4 Local Science Drivers — the Next 2-5 Years

For simplicity, we group discussion of local and remote science drivers in the next section.

11.5 Remote Science Drivers — the Next 2-5 Years

11.5.1 Instruments and Facilities

Data volumes are expanding rapidly — faster than wide-area network bandwidth — in many BER-relevant fields due to improvements in sensors and reductions in storage

costs. These trends are well documented in other scenarios prepared for this requirements review. We should expect these trends to continue and most likely accelerate over the next two to five years.

11.5.2 Process of Science

These developments will increase pressure in three areas of DOE science infrastructure: networks and data transfer tools; storage services; and computational facilities suitable for large-scale data analysis. See below for more discussion.

11.6 Beyond 5 Years — Future Needs and Scientific Direction

This section intentionally left blank.

11.7 Network and Data Architecture

The use cases described above share a common need for reliable, secure, and high-performance end-to-end file transfer. The words "end-to-end" are critical: It serves no purpose to provide a high-speed network into a campus if other factors (e.g., local network configuration, file system configuration) prevent users from making good use of the high-speed wide area connection.

The ESnet Science DMZ concept[5] has proved transformative within the DOE laboratory system as a means of accelerating end-to-end file transfers. It is now quite commonplace for researchers to achieve performance within a factor of two of line speeds for wide area file transfers. Given that prior to the Science DMZ work, performance was often 1 or 2 orders of magnitude less, this is a remarkable achievement.

Looking forward, two major challenges present themselves (see below for more discussion) at the network/data system architecture level: achieving the benefits of the Science DMZ architecture on a much larger scale — extending, for example, to all relevant systems within DOE laboratories, and to many more university campuses; and designing and deploying the storage and computer systems that will be required to meet rapidly growing needs for data storage and analysis.

11.8 Data, Workflow, Middleware Tools, and Services

Previous ESnet reports have clearly shown that file movement and sharing have historically been inordinately difficult, error-prone, and slow, and that such problems collectively represent a major obstacle to the effective use of DOE facilities and to DOE science more broadly.

For example: "Transfers often take longer than expected based on available network capacities"[6]; "lack of an easy to use interface to some of the high-performance tools"[7]; "tools like GridFTP [are] too difficult to install and use"[8]; "high-performance data transfer tools also run into problems with firewalls"[8]; "the effectiveness of data transfer middleware was not just on the transfer speed, but also the time and interruption to

other work required to supervise and check on the success of large data transfers"[6]; "users will do the thing that is easy for them to do, even if it might perform less well than some other more complex solution"[9]; "predictability and reliability are often more important than performance in the realm of data transfer tools"[7]; "facility users do not have the knowledge to troubleshoot data transfers at their home institution"[10]; "the BES community has a need, shared with many other communities, for data transfer tools that are easy to use, well-supported, and permitted by the cybersecurity organizations at the sites and facilities"[8]. These quotations are from reports covering a range of application domains, not just BER, but they all apply to BER science as well.

In summary, major difficulties include:

- Poor end-to-end performance when using commonly available tools, such as secure copy (SCP) — performance that is often only a small fraction of the wide area performance provided
- Difficulty in determining the reasons for poor performance.
- Difficulty in configuring more powerful tools such as GridFTP, for example from the perspective of security configuration and software installation
- Difficulty in dealing with network configuration problems such as firewalls
- Difficulty in managing large transfers, for example detecting and responding to transient network failures, data corruption, and other errors

In response to these problems, a team at Argonne National Laboratory and the University of Chicago has developed, deployed, and operated Globus Online (www.globusonline.org) [11,12], a software-as-a-service file movement solution with convenient Web 2.0 interfaces. In brief:

- Globus Online is a hosted service to which users can direct requests to transfer or synchronize files and directories between two locations. Globus Online handles security, transfer monitoring, and restarts upon failure. It thus automates many of the problematic issues listed above.
- Like SaaS solutions in the consumer and business spaces, Globus Online provides an intuitive Web 2.0 interface for interactive use plus REST interfaces for integration into user applications. A command line interface is useful for scripting.
- Under the covers, Globus Online drives GridFTP transfers and can thus take advantage of Globus GridFTP's highly optimized implementation and ESnet's tuned deployments within Science DMZs.
- Convenient packaging makes deploying local Globus Connect agents on user workstations and laptops and campus clusters a breeze[13].

Since this service was first introduced in late 2010, usage has grown dramatically. As of November 2012, more than 7,000 registered users have moved more than 8 PB and 150 million files. More than 200 endpoints are registered, and major DOE facilities such as ALCF, the Advanced Photon Source (APS), ESG, and NERSC recommend Globus Online to their user communities. ESG has integrated Globus Online as a data download mechanism.

Our experience with Globus Online leads us to draw the following conclusions:

- The Web 2.0 and SaaS approach that has proved so successful in consumer and commercial products can also be used to improve dramatically the quality and usability of research data management capabilities provided to scientists.
- There is a strong synergy between ESnet's work on Science DMZ and Globus Online services. Science DMZ deployments accelerate Globus Online transfers; Globus Online's user-facing capabilities make it trivial for end users to take advantage of Science DMZs.

In recent work, we have developed two additional capabilities that are directly relevant to BER science. We describe them briefly here:

- File-sharing capabilities (demonstrated at the 2012 Supercomputing conference, to be made generally available in early 2012) allow users to manage sharing of files and directories at Globus Online endpoints. Thus it becomes trivial, for example, for a genome scientist to make a new genome data accessible to a set of collaborators.
- Globus Nexus is the identity and group management service that Globus Online uses for such purposes as file sharing. It is trivial for users to define a group, manage group membership, and then allow members of a group to access a shared file. KBase plans to use Globus Nexus for identity and group management.

11.9 Outstanding Issues

There are a number of important outstanding issues related to file movement and sharing, and steps that can be taken to address them.

Performance acceleration. Perhaps the biggest obstacle to really effective use of high-speed networks for science is a lack of knowledge among end users concerning what is possible and among system administrators about what can be done to achieve high performance. An end user who moves data between ALCF and NERSC using Globus Online will achieve high performance, because of appropriately configured DTNs at ALCF and NERSC and because Globus Online will organize the transfer to make good use of those DTNs. But other users may use SCP over the same link, or move data to a university site with poorly configured networks, or not move data at all because they do not realize what can be done[14].

To address this problem, we recommend the following actions:

- Education and outreach programs aimed at end users and system administrators, to inform them of what modern networks can do and how easy it is to make efficient use of those networks
- DOE facilities and ESnet should work proactively to identify slow flows and suggest to end users how to accelerate them. For example, simple analysis of system logs can identify SCP flows. Globus Online can inform end users that their transfers are performing below average relative to their peers, perhaps motivating them to pursue improvements.

- We should provide end users and system administrators with tools that allow them to identify slow flows themselves, and determine possible causes. (See next item.)
- We should investigate the question of where circuit switching (OSCARS, OpenFlow, etc.) can be used to improve end-to-end performance in production settings[15]. Integration with Globus Online is an obvious way to address this question.

Problem diagnosis. An end user may be aware that file transfers are performing badly or unreliably. But it can be extremely difficult to determine the reason for that poor performance. Existing network performance tools are designed for network engineers, not end users. And they only address one part of the problem (e.g., perfSONAR host to perfSONAR host), not the entire end-to-end problem, which may involve local area networks, file system configurations, etc.

To address this problem, we recommend the following action:

- Develop user-oriented performance diagnosis tools that an end user can employ to measure performance, benchmark performance against that achieved by peers, and determine possible reasons for poor performance. Ideally, these tools should be designed to facilitate actions designed to address performance problems. For example, they might say "your performance is only 100 MB/sec, and the reason appears to be the network configuration at location X. Point your system administrator at this report for why we think this is the case, and at this URL for a tutorial on how to correct the problem."
- Integrate these tools into Globus Online so that they can employed easily (even automatically) by end users, without installation of software.

Storage services. Needs are increasingly rapidly for short-term and long-term storage capable of supporting a range of access patterns, from archival (accessed rarely) to online (frequent accesses) and "active" (allowing large-scale computation over data). These services do not exist on anything like the required scale at DOE laboratories. This problem is probably not ESnet's, but doing it right will be important, and how it is done will have important implications for local and wide area network architectures.

Sustainability of services. Globus Online is operated by Argonne National Laboratory and the University of Chicago as a nonprofit service for the research community. As usage grows, so, too, do support and operations costs and demands for expanded capabilities. ESnet may want to contribute to the business-development activities required to achieve sustainability.

11.10 References

1 Williams, D. N. *et al.* The Earth System Grid: Enabling Access to Multi-Model Climate Simulation Data. *Bulletin of the American Meteorological Society* **90**, 195-205 (2009).

2 Johnston, W. Terabit Networks for Extreme-Scale Science. (DOE Office of Science, 2011).

3 Mullender, S. J. & Tanenbaum, A. S. Immediate files. *Software: Practice and Experience* **14**, 365-368, doi:10.1002/spe.4380140407 (1984).

4 Tanenbaum, A. S., Herder, J. N. & Bos, H. File size distribution on UNIX systems: then and now. *SIGOPS Oper. Syst. Rev.* **40**, 100-104, doi:10.1145/1113361.1113364 (2006).

5 *Science DMZ: A Scalable Network Design Model for Optimizing Science Data Transfers*, <http://fasterdata.es.net/science-dmz/> (

6 Dart, E. & Tierney, B. BER Science Network Requirements: Report of the Biological and Environmental Sciences Network Requirements Workshop. (Lawrence Berkeley National Laboratory report LBNL-4089E, 2011).

7 Dart, E. & Tierney, B. FES Science Network Requirements: Report of the Fusion Energy Sciences Network Requirements Workshop. (Lawrence Berkeley National Laboratory report LBNL-644E, 2008).

8 Dart, E. & Tierney, B. BES Science Network Requirements: Report of the Basic Energy Sciences Network Requirements Workshop. (Lawrence Berkeley National Laboratory Report LBNL-4363E, 2010).

9 Dart, E. & Tierney, B. ASCR Science Network Requirements: Office of Advanced Scientific Computing Research Network Requirements Workshop. (Lawrence Berkeley National Laboratory report LBNL-2495E, 2009).

10 Dart, E. & Tierney, B. BES Science Network Requirements: Report of the Basic Energy Sciences Network Requirements Workshop. (Lawrence Berkeley National Laboratory report LBNL/PUB-981, 2007).

11 Foster, I. Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*, 70-73 (2011).

12 Allen, B. *et al.* Software as a Service for Data Scientists. *Communications of the ACM* **55**, 81-88 (2012).

13 Foster, I. *et al.* in *XSEDE 2012* (Chicago, IL, USA, 2012).

14 Kettimuthu, R. *et al.* in *19th IEEE International Symposium on High Performance Distributed Computing* (2010).

15 Liu, Z. *et al.* in *SC12: The International Conference on High Performance Computing, Networking, Storage and Analysis* (2012).

12 ESGF Needs and Strengths from a NOAA Perspective

12.1 Background

Climate science has become networked Big Science. There is considerable agreement on the current set of grand challenges, though not necessarily convergence on a single set of solutions. The recent National Research Council (NRC) report *A National Strategy for Advancing Climate Modeling* identifies the need to foster a diversity of models and approaches united by common scientific goals and infrastructure. This is achieved by model intercomparison projects (MIPs), of which the recently concluded Coupled Model Intercomparison Project (CMIP5) is the largest and most influential. A common set of experiments is agreed upon by all the major modeling centers of the world. This coordinated set of experiments produces data using common standards and protocols, allowing extremely detailed comparisons of different approaches to representing key climate processes. Since the adoption of the MIP approach, there has been an explosion of cross-model studies: The CMIP3 archive resulted in more than 1,000 publications; the CMIP5 archive is on track to surpass that by a wide margin.

The NRC report referenced above recognizes global data infrastructure, and the Earth System Grid Federation (ESGF) in particular, as an essential enabling infrastructure. It further identifies risks in the absence of sustained cross-institutional support for ESGF.

The World Meteorological Organization's (WMO's) Working Group on Coupled Modeling (WGCM) also has recognized the critical role played by ESGF, and talked about its strengths and issues to be addressed in its future development.

12.2 Key Local Science Drivers

12.2.1 Instruments and Facilities

Climate modeling centers participating in CMIP5 operate ESGF *peer nodes*. Currently, 59 models from 24 centers are represented in the ESGF network. Total volume is 1.7 PB.

This represents a substantial growth over CMIP3, which totaled 40 TB six years ago. At that point, a single center, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), was able to host the entire archive (also replicated at the WMO Data Distribution Centers [DDCs]). For CMIP5, a federated system was essential.

NOAA/ Geophysical Fluid Dynamics Laboratory (GFDL) operates an ESGF peer node hosting about 200 TB of model output from four modeling streams: the comprehensive climate model CM3, Earth System Models ESM2M and ESM2G, near-term prediction model CM2.1, and high-resolution models HiRAM C180 and C360.

Figure 23. NOAA's network architecture

N-Wave is currently supporting NOAA's Research and Development High Performance Computing System (RDHPCS) program; portions of NOAA's research, satellite, and ocean services line offices; and is expanding its role in supporting many other mission areas within the agency. NOAA is moving to a Trusted Internet Connection (TIC) architecture with primary access points in Silver Spring, Maryland; Boulder, Colorado; Seattle, Washington; Fort Worth, Texas; and Honolulu, Hawaii.

NOAA's ESGF nodes are at the Princeton and Asheville locations of N-Wave.

12.2.2 Process of Science

To describe the process, consider the following use-case, which is network and compute-intensive.

The study would like to project tropical cyclones in a warming world, with a view to informing policy decisions to prepare cities and coasts for future storms. To arrive at this, resolutions need to become much higher. Current median model resolutions for CMIP5-type climate-change studies are about 100 km; they will need to become about 25 km, representing a 16X increase in data volume over CMIP5.

Understanding which processes contribute to changes in intensity, duration, wind strength, and precipitation efficiency require running detection and attribution studies, where various processes are turned on and off. These types of ensembles are known as

perturbed physics ensembles (PPEs). A brute-force sampling of the parameter space could easily involve thousands of ensemble members; we would take a more parsimonious approach.

To understand the sensitivity of the results to initial conditions, we would run initial-condition ensembles (ICEs). Current studies using high-resolution GFDL models and advanced initialization and data-assimilation techniques show that the size of an ICE also scales wih resolution; more ensemble members are needed as you go up in resolution, though the slope of that power law is not yet known.

Finally, the computation of the key metric (counting storms using physical characteristics, such as vertical temperature structure and vorticity thresholds) is a computationally intensive operation that scales with the horizontal grid size (square of the resolution).

A comparative study across many models would require this analysis to be performed across the federation, touching remote data archives.

12.3 Key Remote Science Drivers

12.3.1 Instruments and Facilities

There are two approaches to performing the analysis in this use case, which have different implications for network technology and software.

One would be for the end user to attempt a systematic download of federated data. Despite the promise of the Science DMZ and the 100 Gbps upgrade of ESnet, this is likely to become unsustainable. Overpeck et al. (2011) estimate the data growth curve in Figure 24.

Despite advances in storage and networks, replicating local stores of federated data is a long-term losing proposition.

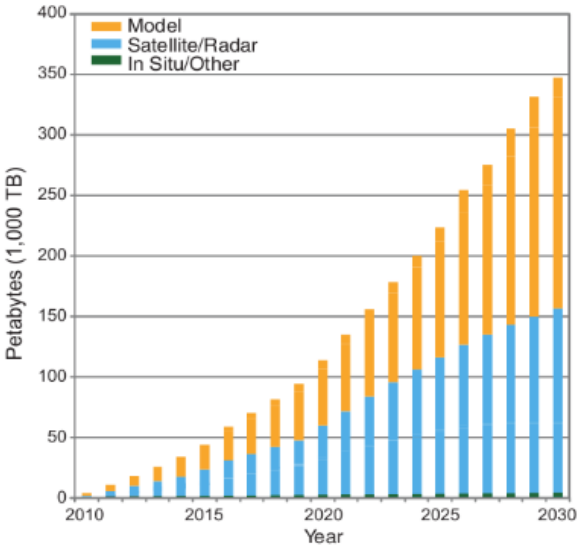


Figure 24. Estimated data growth curve

The second approach is to build server-side analysis and processing capabilities into ESGF. Among various efforts, the ExArch project (funded under an international cross-agency solicitation, the G8 initiative, with the National Science Foundation [NSF] the U.S. agency) is attempting to prototype and scope a system that allows such an analysis to take place across federated archives. These will alleviate network pressure, but will require analysis services (perhaps dedicated clouds) next to the archives. Network latencies may become an issue rather than bandwidth and capacity.

12.3.2 Process of Science

In ExArch prototypes, we have demonstrated the following science process:

- A user queries ESGF gateways, using its faceted search capabilities to locate data sets of interest;
- Generates a list of data sets upon which the server-side “tropical storm counter” analysis package can be run, and picks the ones to use in the current study;
- Dispatches the Open-source Project for a Network Data Access Protocol (OpeNDAP) requests for the required data;
- In its current format, the analysis must run on the ESGF node on which it was launched, and the only remote operations are those supported by OPeNDAP; but the target architecture of ExArch would those also to become remote operations;
- Results are displayed using the Live Access Server layer of ESGF; the data subsets and results can now be downloaded.

12.4 Local Science Drivers — the Next 2-5 Years

12.4.1 Instruments and Facilities

NOAA's key computational facilities are the 1.1 PF Gaea machine located at ORNL and the 0.4 PF Zeus machine located in Fairmont, West Virginia. In addition, NOAA has a substantial allocation on ORNL's Titan machine. NOAA/GFDL additionally has been awarded an Early Science Project on Argonne's Blue Gene/Q (BG/Q) of 150 million CPU hours. NOAA/GFDL hosts a 40 PB archive with associated analysis and storage facilities at Princeton, New Jersey.

12.4.2 Process of Science

NOAA/GFDL runs an integrated workflow management system, the FMS Runtime Environment (FRE). It manages long-running climate simulations in the queues of all machines such as Gaea, Zeus, and Titan; manages the data transfers back to the Princeton analysis facility; and executes post-processing and analysis at the Princeton facility, all the way to publication into ESGF for CMIP5 and other projects.

Data integrity is a key concern. GridFTP is the principal mechanism for data transfer. Several layers for data integrity and caching are built into the FRE system; we would like to cede those operations to gridFTP and Globus as they become available.

The data traffic between computing sites and the Princeton analysis facility is managed over dual redundant 10 Gbps N-Wave links. Out of a peak capacity of 80 TB/day, we are currently experiencing sustained throughput of 24 TB/day.

Network latencies are also a concern for keystroke traffic to remote sites. These are currently managed using open-source network accelerators such as FreeNX.

12.5 Remote Science Drivers — the Next 2-5 Years

12.5.1 Instruments and Facilities

12.5.2 Process of Science

12.6 Beyond 5 Years — Future Needs and Scientific Direction

12.7 Network and Data Architecture

NOAA has a significant intererest in the Science DMZ (http://fasterdata.es.net/science-dmz/) and the federal Big Data Initiative. We plan to continue to engage with DOE in fostering these interactions.

12.8 Collaboration tools

Remote collaboration and reduction of travel continue to be NOAA concerns. There are no special requirements beyond those of other agencies.

12.9 Data, Workflow, Middleware Tools, and Services

We have attempted throughout this document to articulate a case for a substantial change in the portfolio: to invest more in workflow and service middleware layers relative to network hardware. This would be our principal recommendation.

Following the language of the NRC report, we would recommend that ESGF be treated as infrastructure (ESnet) rather than a software research project (Scientific Discovery through Advanced Computing [SciDAC]). This would imply that ESnet should seek cross-agency agreements with NOAA and other agencies (including international partners) to place ESGF under community technical governance and infrastructure funding.

13 Banfield Lab Omics Workflows at UC Berkeley

13.1 Background

Our group at the University of California (UC) Berkeley conducts research on microbial ecosystems using high-throughput molecular biology methods and bioinformatics. These primarily center on metagenomics (reconstructing genomes from environmental samples as opposed to laboratory cultures), proteomics (assessing the proteins present in a sample), transcriptomics (assessing the actively created RNA messages in a sample), and metabolomics (understanding the diversity of metabolites present in a sample). Integrating these various data streams is critical to our scientific process.

13.2 Key Local Science Drivers

13.2.1 Instruments and Facilities

Our group is primarily an end user of many of the national laboratories. As such, our instruments and facilities revolve around computational hardware to assist us in analysis and integration of these data streams.

Our current compute infrastructure consists of the following:

- Assembly servers (2X): used for assembling metagenomes; 32/40-core Intel Xeon processors, 1024 Gb system memory
- Annotation server: used for analyzing gene/protein function; 24-core Intel Xeon processors, 128 Gb system memory
- Integration server: used for supporting our KBase database; 24-core Intel Xeon processors, 128 Gb system memory
- Web server: used for supporting our KBase presence; 8-core Intel Xeon processors, 24 Gb system memory
- Data storage SANs (3X): two 80 Tb and one 40 Tb with 8 Gbps Fibre Channel connections to servers

This hardware is housed in the UC Berkeley data center, a climate-controlled and secure facility. Connectivity at the facility includes redundant Gbit-fibre connections to the Internet.

Communication with our source data streams occurs via standard protocols (HTTP/S, FTP, SCP/FTP).

13.2.2 Process of Science

Our main goal with metagenomics is to create as many complete or near-complete genome sequences as possible for as many different representatives of the community being studied as possible. This last point is critical: For organisms in very low abundance, extremely large amounts of sequencing data are required for adequate assembly.

After sample collection and processing, our data workflow begins with an assembly of the DNA sequence for a project. Typically, we end up with about 350 million reads per

lane of sequence. These numbers can vary, depending on amount and quality of the DNA preparations. Sequencing read length is currently running 100 to 150 base pairs per read. This results in 35 to 50 Gbp per lane and typical projects are currently generating anywhere from 3 to 15 lanes of sequence. Although the raw sequence data are heavily processed later in the workflow, it is still critical data and must be archived. We keep archives locally as well as at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA).

We perform assemblies of the raw (or quality-trimmed) reads in order to form larger contiguous sequences (contigs). Several short-read assembly software tools are available. We spend a great deal of time in assembly refinement and have developed several protocols for maximizing assembly output. Unfortunately, in many situations additional custom workflows must be developed (there is no one-size-fits-all workflow for assembly, in our opinion).

Once we have an assembly, the next step is binning. This involves many processes and results in the formation of a bin: a collection of contigs that belong together and likely originate from a single host organism (or a set of very closely related organisms or strains). Binning procedures range from looking at detailed sequence statistical analysis (e.g., self-organizing maps of di-, tri-, tetranucleotide frequencies) to looking at more broad phylogenetic labels we generate during our QuickLook protocol. The QuickLook collects a set of statistics about each contig, including GC%, codon usage and genetic code, read depth or coverage, time-series abundance data, and overall phylogenetic classification. These data are critical for comprehensive binning. The predictions from the QuickLook are then integrated with the statistical sensitivity of the self-organizing map procedure.

When we feel the assembly and binning has reached a high quality, we freeze the bins and do a thorough functional annotation of the predicted genes for every bin. This process is time-consuming (CPU-bound) and includes both sequence similarity searches (using USEARCH instead of Blast) as well as motif-level pattern searches (using InterproScan). The goal of this step is to develop as much information as possible about the genes present in a bin. Using the functional annotation for a bin, we determine how complete it is, what its metabolic capabilities are, and how variable its genome is.

We have developed a system with funding from the DOE KBase initiative (S. Gregurick, Program Manager) called ggKBase. This is a Web service and application programming interface that integrates microbial data. After a bin is annotated, everything about the gene — sequence, location, statistics, annotations — is incorporated into ggKBase for further work. The tools in ggKBase allow further bin refinement and validation, cross-bin comparisons for sequence and function, a complex and powerful searching engine, and a comprehensive and social list-building system. These tools allow the user to answer intricate questions about an organism (or group of organisms) such as what are the metabolic pathways present or absent and how do they compare with other organisms in the project or in other projects? Finally, ggKBase contains (via RESTful service calls) information from other data streams — metabolites, protein levels, or transcript levels

— for any sample that has been analyzed with one of our collaborators at the national laboratories.

13.3 Key Remote Science Drivers

13.3.1 Instruments and Facilities

We receive data from the Joint Genome Institute (JGI) (sequence), Oak Ridge National Laboratory (ORNL) and Pacific Northwest National Laboratory (PNNL) (proteomics), and Lawrence Berkeley National Laboratory (LBNL) (metabolomics). Additionally, we interact with NCBI to deposit our genomics data at GenBank. All data transfer is done via standard protocols mentioned above.

13.3.2 Process of Science

As end users, we have little to do with the remote operations with our data stream sources at the national labs. Our main involvement is with sample exchange, handling discussions, and data and protocol exchanges. Our goal with our collaborators is largely quality scientific publications and the education of next-generation scientists.

13.4 Local Science Drivers — the Next 2-5 Years

13.4.1 Instruments and Facilities

We anticipate a somewhat linear increase in our local computer hardware over the next two to five years. We have been adding approximately one well-configured server to our system every year for the past for five years. The expense with these additions comes with the significant amounts of system memory required for assembly software. Additionally, we have been adding approximately 40 TB of SAN storage to our system each year. We anticipate this will grow, possibly by 100% in the next two- to five-year range.

13.4.2 Process of Science

The process of science will remain what it currently is and adapt to new software tools that we (or others) develop.

13.5 Remote Science Drivers — the Next 2-5 Years

13.5.1 Instruments and Facilities

(same as above)

13.5.2 Process of Science

(same as above)

13.6 Beyond 5 Years — Future Needs and Scientific Direction

In omics and environmental research, five years is a very long time. On the immediate horizon are long-read, high-quality sequencing technologies that will make many of the nuts-and-bolts tasks of current omics research redundant. We anticipate an era when all metagenomics studies yield hundreds, probably thousands (or more), complete (or very nearly so) genomes from complex time series/experimental data sets. Thus, very soon, we expect that the primary focus of research will be data analysis, not bioinformatics as currently viewed. If this future arrives, research needs will center on genome analysis and functional data-integration tools that can enable massively parallel biological/ecological analyses. For this reason, ggKBase has been developed around the concept of simultaneous analysis of metabolism and function in as many genomes as are present within the analysis set.

13.7 Network and Data Architecture

GgKBase is a data-integration tool that leverages RESTful Web services. One potential issue is the lag that can occur, for example, when requesting metainformation for 10,000 metabolites generated from a metabolomics run for cross-referencing to a metagenomic sample. Having a very fast network would facilitate this interaction and accelerate knowledge discovery. Other than this, our work has no special requirement for high-performance data transfers or specialized network hardware relative to other Big Data applications.

13.8 Collaboration Tools

We have no special needs for collaboration tools: We primarily use e-mail and Skype to interact with our collaborators.

13.9 Data, Workflow, Middleware Tools, and Services

As the amount of DNA sequence increases, significant barriers will occur with our existing pipeline for metagenomic analysis. For one, sequence assembly requires a significant amount of system memory. We are already hitting this barrier even on our 1 TB servers. Complex samples are therefore limited to approximately one lane/350 million reads per assembly (we have assembled up to 850 million reads for a less-complex experiment). This will prevent us from doing combination assemblies across lanes. Additionally, assembly is a computation bottleneck. Advancements in using cloud computing for assembly have been made recently, for example Contrail, but these efforts are very new and the resulting applications untested. However, as mentioned above, assembly strategies often need to be customized and benefit from having multiple assemblies for comparison. Software developers need the tools to take advantage of scaling their apps onto cloud-computing infrastructure.

13.10 Outstanding Issues

The amount of data generated in *omics projects is massive and growing fast. We use a backup strategy that basically keeps two copies of all the original data, plus a copy is deposited at NCBI in the SRA. Locally, this amounts to a considerable amount of storage space. Data compression for DNA sequence data is an important area that could be further developed. Simple compression (zip/gzip) works fine, but significant advancements could be leveraged with “lossy” data compression that may help to increase the compression ratio.

Although bioinformatics has made significant advances in methodology and application recently, it still suffers from a lack of communication between groups. This is a challenging problem because the field is too new to benefit from rigorous standards and the field as a whole is changing quickly. Clear connections between groups with easier communication and protocol exchange could greatly enhance bioinformatics research.

Further, metagenomics is an international undertaking. In fact, some of the most important breakthroughs are coming from groups outside of the United States, e.g. (based on the August 2012 International Society for Microbial Ecology [ISME] meeting), Denmark and Per Nielsen's group and Australia (Gene Tyson's group). The methods are evolving fast, as they must to keep up with breathtakingly rapid improvements in technology (improvements in data size, quality, and type) and expanding topic areas (e.g., human microbiome time course experiments, studies of the response of ecosystems to increased CO_2 levels, in situ evolution studies, etc.). Consequently, any attempt to standardize protocols, formats, or procedures can only hinder progress in the field and may not be accepted in the broader scientific community.

13.11 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• Multiserver configuration and data storage	• Assembly/annotation of 30-50 Gbp of sequence data	• 10-50 TB • Variable data file sizes and quantities	• 100 GB approx. 10/day	• 1 TB/week • Data transfer between national lab collaborators
2-5 years				
• Additional servers and data storage. • Beginning move to cloud-based strategies	• Adaptations to workflows to support additional compute resources and cloud infrastructure	• Similar to original breakdown	• 100 GB approx. 10/day	• 5 TB/week. • Same as above
5+ years				
• Same as 2-5 year	• Same as 2-5 year	• Increase of 50 to 100% over 2-5 year period • Similar to original breakdown	• Similar to above	• Similar to above

14 EMSL

14.1 Background

The Environmental Molecular Sciences Laboratory (EMSL) is a national user facility that provides world-class fundamental research capabilities for scientific discovery and the development of innovative solutions to the nation's environmental challenges and energy production. EMSL's distinctive focus on integrating computational and experimental capabilities as well as collaborating among disciplines yields a strong, synergistic scientific environment. Bringing together experts and state-of-the-art instruments critical to their research under one roof, EMSL has helped thousands of researchers use a multidisciplinary, collaborative approach to solve some of the most important national challenges in energy, environmental sciences, and human health. These challenges cover a wide range of research, including synthesis, characterization, theory and modeling, dynamical properties, and environmental testing.

EMSL is located in Richland, Washington, and is operated by PNNL for the DOE Office of Biological and Environmental Research.

14.2 Key Local Science Drivers

14.2.1 Instruments and Facilities

EMSL houses an unparalleled collection of state-of-the-art capabilities that are used to address scientific challenges relevant to DOE and the nation. Researchers from around the world are encouraged to use EMSL's unique capabilities in combination with one another with an emphasis on merging computational and experimental instruments. EMSL is currently developing MyEMSL, a framework for scientific collaboration based on data from both internal and external sources.

EMSL consists of multiple experimental capabilities. Each EMSL capability operates a set of scientific instruments on behalf of EMSL users. The capabilities include:

- Cell Isolation and Systems Analysis
- Deposition and Microfabrication
- Mass Spectrometry
- Microscopy
- Molecular Science Computing
- Nuclear Magnetic Resonance and Electron Paramagnetic Resonance
- Spectroscopy and Diffraction
- Subsurface Flow and Transport

Capabilities with significant networking needs are described in more detail below.

Cell Isolation and Systems Analysis is used to isolate cells from complex populations or environmental samples for subsequent integrated omics and imaging analyses. EMSL specializes in high-throughput genomics and proteomics studies as well as electron and fluorescence microscopy characterization at high spatial and temporal resolutions.

These capabilities provide the foundation for attaining a molecular-level understanding of protein-network and microbial-community dynamics and enable the pursuit of systems biology, including the new field of systems microbiology. Instruments include cell isolation and fractionation resources, fluorescence microscopes and spectroscopes, electron microscopes, and transcriptomics instruments to perform massively parallel next-generation sequencing.

Mass Spectrometry enables high-throughput, high-resolution analysis of complex mixtures. These resources are applied to a broad range of scientific problems, from proteomics studies with applications to human health and environmental remediation to aerosol particle characterization, as well as fundamental studies of ion-surface collisions and preparatory mass spectrometry using ion soft-landing. Instruments include Fourier transform (FT) mass spectrometers, including FT ion cyclotron resonance (FTICRs), Orbitraps and linear trap quadrupole (LTQ)-Orbitraps, linear ion traps, triple-quadrupole spectrometers, ion-mobility spectrometry (IMS), time-of-flight (TOF) spectrometers, high-performance liquid chromatrography (HPLC), a field-deployable second-generation single-particle laser-ablation TOF mass spectrometer, and an ion soft-landing deposition instrument.

Microscopy has a wide variety of sophisticated microscopy instruments, including electron microscopes, optical microscopes, scanning probe microscopes, and computer-controlled microscopes for automated particle analysis. These tools are used to image a range of sample types with nanoscale — and even atomic — resolution with applications to surface, environmental, biogeochemical, atmospheric, and biological science. Each state-of-the-art instrument and customized capability is equipped with features for specific applications. Instruments include electron microscopes with tomography, cryo, scanning, photoemission, and high-resolution capabilities; a nuclear magnetic resonance (NMR) microscope; a dual Raman confocal microscope; optical microscopes; single-molecule fluorescence tools; spectroscopy tools with visible, near-, mid-, and far-infrared capabilities; atomic-force microscopy; and scanning probe microscopes.

Molecular Science Computing provides an integrated production computing environment supporting a wide range of computational activities in environmental molecular research, archive storage, scientific expertise, and the NWChem computational chemistry software suite. Systems include a 2310 node supercomputer (with a peak of 163 Tflops, consisting of dual quad-core AMD Opteron processors, 37 TB of memory, a 300 TB global file system), a 6 PB hierarchical archive storage system named Aurora, and a graphics and visualization system.

Spectroscopy and Diffraction has a suite of spectroscopy and diffraction instruments in EMSL, allowing users to study solid-, liquid-, and gas-phase sample structure and composition with remarkable resolution. Ideal for integrated studies, spectrometers and diffractometers are easily coupled with EMSL's computational and modeling capabilities, enabling users to apply a multifaceted research approach for experimental data interpretation and to gain a fundamental understanding of scientific problems.

Instruments include electron spectrometers; various electron microscopes; FT infrared spectrometers; an ion accelerator system; five Mössbauer spectroscopy systems; optical spectroscopy tools including confocal-Raman, time-resolved fluorescence, and second harmonic generation capabilities; and multiple X-ray diffraction instruments.

EMSL's internal network core is built around a fiber infrastructure that provides connectivity to a standard eight-port network switch in each office and laboratory space. The capability exists to directly attach gigabit and 10 gigabit instrumentation and computational resources directly to the EMSL core network. Isolated instrumentation networks are created using the building fiber to interconnect lab spaces throughout the EMSL building, and in some cases extending into other buildings. The EMSL core network has multiple connections into the PNNL core network through redundant fiber paths.

14.2.2 Process of Science

EMSL's capabilities are available to researchers through a peer-reviewed proposal process, at no cost, if research results are published in the open literature. Users access the facility to use one or more capabilities, and work with EMSL’s expert staff to gain insight and knowledge into their scientific problem. A large majority of instruments at EMSL require hands-on work and the assistance of scientific experts from EMSL.

Data are generated by most instruments, and usually processed either automatically or manually before delivery to a user. It is shipped to the user’s home institution through e-mail, and by media such as CD and thumb drives when necessary due to bandwidth limitations. In extreme cases, hard drives are shipped to the user’s home institution, owing to the quantity of data and the uncertainty of reasonable bandwidth between EMSL and the home institution.

The Aurora storage archive at EMSL is increasingly being used as a central store for EMSL data. It contains 4.5 PB of user data. Presently, EMSL produces about 45 TB of data weekly.

14.3 Key Remote Science Drivers

14.3.1 Instruments and Facilities

EMSL’s users have remote access to the Chinook supercomputer and the Aurora data storage archive using remote tools to access some instruments, saving time and travel costs.

EMSL has multiple and redundant connections to ESnet (and the Internet) via 10 Gbps links through PNNL to Seattle (primary) and Boise (failover). The network was engineered with reliability and performance in mind: Should one of the 10 Gbps links suffer an incident that disrupts primary service, traffic is automatically failed-over to the redundant link.

14.3.2 Process of Science

Remote data sources are increasingly important as EMSL develops more of a focus on team-based science and integration of data from multiple sources. We expect much of this to be driven by systems biology, where data produced by other institutions will be transported to EMSL for integration, analysis, and visualization. Likely remote endpoints in this scenario include the forthcoming Biosystems Frontier Facility, Joint Genome Institute (JGI), the Joint BioEnergy Institute (JBEI), and the Systems Biology Knowledgebase (KBase). EMSL and JGI have already established new interfaces for automated download of data from JGI to EMSL. MyEMSL should become a focal point for data transfer and collaboration activities by EMSL users and their collaborators.

14.4 Local Science Drivers — the Next 2-5 Years

14.4.1 Instruments and Facilities

New generations of ultrafast and high-resolution electron microscopes will drive new data growth. The first of the new generation (ultrafast transmission electron microscopy [UTEM]) will produce a peak of 11-13 TB/day, in batches of approximately 700 files.

EMSL expects to take delivery of the new HPCS-4A high-performance computing system in the summer of 2013. The new system will provide over 3 petaflop/s of computing capacity, and have a parallel file system in excess of 2 PB capacity. The system is expected to be used mostly for computational chemistry and climate modeling. This system will likely produce 1-2 PB of climate data, and possibly hundreds of TB worth of molecular dynamics trajectory files. In addition to scientific computing, the supercomputer will increasingly be used to provide real-time analysis of the experimental data streaming off the scientific instruments. In addition, in the next five years the archive storage system will be upgraded to reach 20 PB of data storage.

EMSL is in the early stages of developing a High Resolution and Mass Accuracy Capability (HRMAC). A next-generation HRMAC is needed to ensure that EMSL will continue to provide leading-edge resources to serve national and international users who are addressing critical DOE mission needs. Next-generation capability, taking advantage and applying newly developed technologies and approaches, will address this need, significantly enhancing overall analytical and characterization performance in terms of sensitivity, dynamic range, accuracy, resolution, and speed or throughput, and enable previously intractable types of applications.

The common denominator is increased resolution and increased data rates coming off the instruments. New data management policies and processes will improve EMSL's ability to make unique data available to the scientific community. This should make EMSL a supplier of PB of data to ESnet's users. It is anticipated that the archived proteomics data will be accessed with increasing frequency as its use in gene annotation becomes more common. Thus, the volume of accessed data should increase by 2- 5-fold in the next several years. Peak bandwidth to allow timely access to remote instrumentation will continue to be a factor.

14.4.2 Process of Science

One change that will occur in this time frame is the increased production use of an EMSL-wide scientific data management system (MyEMSL) that stores data acquired from EMSL's experimental and computational instruments and the output of analysis software, together with relevant metadata. MyEMSL will provide users a simple way to find, retrieve, visualize, and analyze their stored data. The data management system will provide a simple and consistent interface for the EMSL staff operating the instruments and for users accessing and sharing their data. MyEMSL will likely be a catalyst for creation of new data and discoveries derived from existing data.

14.5 Remote Science Drivers — the Next 2-5 Years

14.5.1 Instruments and Facilities

In the two- to five-year time frame, the next-generation supercomputing capability will become available to the EMSL user community. Increased size and number of raw data sets will make it more difficult for users to move the data to their home institutions for analysis, which drives the need for increased access to remote analyses and visualization capabilities at EMSL.

14.5.2 Process of Science

EMSL expects the combination of MyEMSL, higher-quality image data, and systems biology to drive more hosting of data sets by EMSL for access by external parties. There will be increased interest in remote collaboration, in which data are posted for shared access, and collaborators can share information about it in real time.

14.6 Beyond 5 Years — Future Needs and Scientific Direction

EMSL plans to increase its scientific impact during its second decade of operation by focusing attention and capability development in specific areas identified as high-priority science themes. These science themes help define and direct development of key capabilities and collections of user projects that can have significant impacts on important areas of environmental molecular science critical to DOE and the nation.

As EMSL's user research expands and matures, new and enhanced capabilities will be developed. Additionally, existing systems will be modified to support the needs of the user community. Beyond five years, we should see growth in the body of biological data that EMSL and PNNL will maintain for searches and analysis. The increased focus on systems biology, and improved imaging capabilities, will significantly increase the data volumes for complex samples analyzed by EMSL. A new generation of mass spectrometers for proteomics applications is being developed that should increase sample throughput and data output by multiple orders of magnitude. Access to the massive sets of data generated by these new instruments will significantly increase network requirements. Much of this data will be transferred in from off site, combined with existing data, curated, and shared back with the scientific community. Strong

integration of data from multiple scientific domains to allow users to address systems-level problems will require EMSL to manage and integrate multi-PB-scale data sets. This will require the development of complex workflows, accessing data generated and stored at EMSL with data from other user facility and research laboratories, which will significantly impact network requirements.

14.7 Network and Data Architecture

No explicit Big Data initiatives have been identified at this time.

PNNL currently has a data transfer node (dtn.pnl.gov) attached to its Secure Collaboration Zone (SCZ) perimeter network. The dtn.pnl.gov system has a 10Gbps connection to the Internet and QDR IB attached to a 4 PB Lustre storage cloud; it supports 1 GB/sec data transfers. The storage cloud has multiple internal mount points, and is available to the Olympus supercomputer via QDR IB interconnects. EMSL also has its Aurora archive attached to the SCZ network at 10Gbps, providing up to 1 GB/sec data transfer capability to other laboratories. The SCZ has a perfSONAR/NDT testing point attached at 10 Gbps (ndt.pnnl.gov). The SCZ utilizes a host-based firewall model in which a Port Scan Attack Detector (PSAD) is used in conjunction with iptables to detect and block attackers with little performance degradation on individual hosts.

PNNL has not started down the path to engineer any DTN on 100 Gbps capability, and more importantly, we have very few data transfers that use more that 100 MB/sec streams. PNNL does not really exercise the existing 10 Gbps capability.

14.8 Collaboration Tools

MyEMSL will provide a set of collaboration tools for users and their collaborators, and will rely on a standard set of protocols: HTTP, SSH, FTP, VNC. There is an increasing interest in Skype, and to a lesser degree, EVO.

14.9 Data, Workflow, Middleware Tools, and Services

EMSL has yet to determine what middleware and services might be required. Some interest has been expressed in limited access to external private and public clouds.

14.10 Outstanding Issues

EMSL frequently needs to ship physical copies of media to users when data sizes exceed a few GB. More often than not, this is due to lack of bandwidth or storage resources at the user's home institution.

14.11 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• Broad suite of scientific instruments • Chinook 160 Tflop supercomputer • Quiet wing with hi-res chemical imaging capabilities • Systems biology	• Primarily on-site access to instruments • Remote access to Chinook computer, Aurora archive, and remote instrument operation • MyEMSL data management system will store data (and metadata) of all instruments	• Data volume 6 TB/day	• 5 TB/day continuous, 24x7	• 200 GB/month at 1 GB/sec • Data transferred to users' home institutions
2-5 years				
• Next-generation HPCS-4 A (3+ Tflop) and B supercomputers • Next-generation mass spectrometer • Next-generation electron microscopy	• Enhanced integration of data from multiple instruments • Collaborative data access and analysis • MyEMSL with search and first set of workflow and analysis capabilities	• Data volume 20 TB/day	• 40 TB/day continuous, 24x7	• 600 TB/month at 1 GB/sec to user's home institutions • 5 TB/month at 1 GB/sec from JGI, JBEI, KBase
5+ years				
• HPCS-5 HPC system(s) • Next-next generation electron microscopy • Enhanced imaging instruments	• Strong integration of data across capabilities • Comprehensive problem-solving environment on top of MyEMSL	• Data volume 100 TB/day	• 200 TB/day continuous, 24x7	• 3 PB/month at 10 GB/sec to user's home institutions • 50 TB/month at 10 GB/sec from JGI, JBEI, KBase

15 Great Lakes Bioenergy Research Center

15.1 Background

Biofuel production has the potential to mitigate climate change and provide energy security. However, important questions about biofuels must be addressed to ensure their production, economic, and environmental efficacies. These questions relate, for example, to production volumes, land requirements, competition with other land uses, mitigation effects, and long-term effects on water and soil resources. The DOE Great Lakes Bioenergy Research Center (GLBRC) has developed a high-resolution, multiscale modeling framework for the study of these questions. The framework consists of (a) the Environmental Policy Integrated Climate (EPIC) model; (b) a 60 m resolution geodatabase of the conterminous United States, containing data on climate, soils, topography, hydrography, and land cover/land use; (c) ancillary data (e.g., road networks, federal and state land, national and state parks, etc.); and (d) algorithms to conduct optimization and uncertainty analyses. The modeling framework has been applied so far at multicounty, large-regional, and subnational scales. Simulation results of biomass productivity and environmental outcomes (soil carbon change, soil erosion, nitrate leaching, and nitrous oxide emissions) are used to develop spatially explicit bioeconomic and Life Cycle Analysis models.

Over the past five years, GLBRC researchers (from PNNL and ORNL) have developed an advanced computing infrastructure to execute millions of biophysical and biogeochemical simulations, conduct post-processing calculations, store input and output data, and visualize results (Zhang et al., 2010). At PNNL's Joint Global Change Research Institute (JGCRI), the computing resources include two components installed at the Research Data Center of the University of Maryland. The first resource is deltac, an eight-core Linux server dedicated to county- and state-level simulations and PostgreSQL database hosting. The second resource is the DOE-JGCRI Evergreen computer cluster, capable of executing millions of biophysical simulations in relatively short periods.

15.2 Key Local Science Drivers

15.2.1 Instruments and Facilities

The Evergreen computing cluster comprises 284 compute nodes, plus a handful of auxiliary nodes. All user jobs are run on the compute nodes, while the auxiliary nodes provide services such as interactive logins, network access to the outside world, and access to long-term storage. Each compute node hosts a pair of Intel Xeon 5560 (quad-core) or 5660 (hex-core) microprocessors running at 2.8 GHz. The system has a total of 2,472 cores available. Each core has a peak throughput (*R*peak) of 11.2 Gflops (billion floating-point operations per second). Thus, the total *R*peak for the system is 27,686 Gflops, or approximately 27.7 Tflops. Each node has 48 GB of memory, for a total system memory of 13.6 TB.

Two user-accessible networks link the compute nodes. The primary network is a QDR InfiniBand (IB) network running at 40 Gbps line rate (32 Gbps data rate). The round-trip latency (i.e., the time required to send a zero-length message and receive a response) on the IB network is 1.7 microseconds. Thus the IB network provides both high throughput and rapid response times. It is therefore ideal for running distributed calculations in which the compute nodes performing the calculations must exchange data frequently.

The backup network runs on Gigabit Ethernet. This network has much lower throughput and much higher latency (typically around 100 microseconds). It is therefore unsuitable for calculations that require extensive communication between compute nodes. In practice, Evergreen's Ethernet network does see some use in calculations that for technical reasons are unable to use the IB network.

Evergreen users' storage needs are supported by a 1.4 PB Lustre file system accessible from all compute nodes and auxiliary nodes in the system. Lustre is designed to allow parallel access to files from any or all compute nodes simultaneously. This capability supports higher read/write performance than is achievable with a conventional network file system. The file system is fully backed up and available for long-term storage for user data.

15.2.2 Process of Science

The Python programming language is used to create high-resolution geodatabases, set up model simulations, and process simulation results. Managing vast amounts of simulation results is done via the freely available PostgreSQL server. Preparing spatial data for input to biophysical models is done using the ArcGIS™ software, using the application programming interfaces (APIs) exposed to the Python programming language.

Model calibration is done on the deltac server (16 computational nodes, with hyper threading enabled and 32 TB storage capacity). Recently, code was developed to conduct multivariable calibration (e.g., plant and soil parameters) on Evergreen. PostgreSQL resides on deltac and can be accessed by users over the Internet. Special databases in Excel format are distributed to agricultural economists to conduct bioeconomic analyses and to bio-engineers to conduct life-cycle analyses. Both deltac and Evergreen are hosted at the Resources Data Center of the University of Maryland, College Park, Maryland.

Using high-performance computing (HPC), collaborators at ORNL completed >140 K simulations in ~10 h on an HPC cluster using 20 nodes; a speedup of 40 times over conventional computing resources (Nichols et al., 2010). Multistate runs (up to 30 states so far) are conducted on Evergreen. In addition to biomass analysis under the GLBRC project, results on agricultural productivity are used as input for the Regional Global Climate Assessment Model (RGCAM).

15.3 Key Remote Science Drivers

15.3.1 Instruments and Facilities

Most large-scale data transfer for the GLBRC project has occurred between ORNL and PNNL's JGCRI. Data transfer occurs between servers similar to deltac.

Besides the GLBRC group, the main user of Evergreen is JGCRI's integrated assessment (IA) group running and developing the Global Change Assessment Model (GCAM), a dynamic-recursive model (economy, energy, and land use) that includes numerous energy-supply technologies, an agriculture and land-use model, and a reduced-form climate model. Members of the IA modeling community access Evergreen regularly to conduct simulations with diverse IA models. Other groups access Evergreen to perform regional Earth systems and climate modeling experiments.

15.3.2 Process of Science

In addition to high-resolution biofuel modeling experiments conducted under the GLRBC program, other scientific experiments are performed such as those related to climate stabilization driven by numerous energy technologies, agriculture, and land use.

15.4 Local Science Drivers — the Next 2-5 Years

15.4.1 Instruments and Facilities

Over the past five years, GLBRC biofuels modeling has evolved from PCs, to PC clusters, to servers, and to supercomputers (at PNNL's JGCRI and ORNL). Computation, storage, and network capabilities are anticipated to increase significantly over the next two to five years as the GLBRC project initiates its second five-year phase in December 2012 (FY 2013–FY 2017).

This increase will be needed to store, transfer, and handle larger input/output data sets than those exercised so far. Currently, storage limitations on servers are met by reducing model output to a minimum. Increases in storage, transfer, and retrieval capabilities will be essential to handle the new experiments anticipated for the GLBRC. Currently, geospatial data consumes about 5 TB of storage space and is expected to double or triple during the second phase of the GLBRC project. Additional storage and handling capabilities will be needed for bioeconomic and life-cycle assessment (LCA) data.

15.4.2 Process of Science

Spatially explicit biophysical (e.g., yield) and biogeochemical (e.g., greenhouse gas production and flow) simulations will continue to dominate computational activities. Biorefinery modeling by interdisciplinary groups will required at least 1 TB per biorefinery modeled. For Phase II, multiple biorefinery-scale simulations and large-scale simulations of the U.S. Midwest and the conterminous United States will produce numerous TB of data.

15.5 Remote Science Drivers — the Next 2-5 Years

15.5.1 Instruments and Facilities

Access and sharing of data should be improved over the next two to five years. Implementation of a methodology like the one being employed in the AgMIP project, where different teams can agree on a common language and platform to exchange data, would be beneficial. Currently, the GLBRC lacks a good way to share, archive, manage, and disseminate large quantities of data. Perhaps the GLBRC wiki already allows this, but it does not seem to be very popular.

The next two to five years will see a need to develop products that can reach the end users (e.g., Web sites, coupled with requisite licenses for ArcGIS online services to serve spatial data online).

15.5.2 Process of Science

The biofuel modeling activities during the next two to five years will be increasingly collaborative and interdisciplinary in nature. Effective communication among scientists and effective information exchange are deemed essential for the success of the research enterprise.

15.6 Outstanding Issues

None.

15.7 References and Additional Material

Gelfand, I., R. Sahajpal, X. Zhang, R.C. Izaurralde, K.L. Gross, and G.P. Robertson. 2013. Sustainable bioenergy production from marginal lands in the US Midwest. Nature 493: doi:5140519. 10.1038/nature11811.

Egbendewe-Mondzozo A., S.M. Swinton, R.C. Izaurralde, D.H. Manowitz, and X. Zhang. 2011. Biomass supply from alternative cellulosic crops and crop residues: A spatially-explicit bioeconomic modeling approach. Biomass Bioenergy doi:10.1016/j.biombioe.2011.09.010.

Nichols, J., S. Kang, W. Post, D. Wang, P. Bandaru, D. Manowitz, X. Zhang, and R.C. Izaurralde. HPC-EPIC for high-resolution simulations of environmental and sustainability assessment. Computers and Electronics in Agriculture 79:112-115.

Zhang, X., R. Srinivasan, J.G. Arnold, R.C. Izaurralde, and D. Bosch. 2011. Simultaneous calibration of surface flow and baseflow simulations: a revisit of the SWAT model calibration framework. Hydrological Processes 25:2313-2320. DOI: 10.1002/hyp.8058.

Zhang, X., R.C. Izaurralde, D. Manowitz, T.O. West, W.M. Post, A.M. Thomson, V.P. Bandaru, J. Nichols, and J.R. Williams. 2010. An integrative modeling framework to evaluate the productivity and sustainability of biofuel crop production systems. Global Change Biol. – Bioenergy 2:258–277.

16 Joint Genome Institute, Walnut Creek, CA

16.1 Background

The Joint Genome Institute (JGI) supports the Department of Energy's mission in the areas of bioenergy, carbon cycling, and biogeochemistry. JGI is DOE's sole production genomic sequencing center. It sequences and annotates genomic data for nearly 1,800 users from a variety of scientific community and DOE projects, including the Bioenergy Research Centers (BRCs). JGI provides access to various genome data sets through a number of portals (Genome Portal, Integrated Microbial Genomes [IMG], Phytozome, etc). As a user facility, JGI also supports the submission and analysis of sequence data from external collaborators and other scientists. These activities result in Web traffic and large data movements both to and from JGI, throughout the United States and internationally.

16.2 Key Local Science Drivers

16.2.1 Instruments and Facilities

JGI operates several sequencing platforms (Illumina and Pacific Biosciences) and requires significant computing resources to process and analyze the data. Currently, the Illumina HiSeq sequencers are capable of generating the largest data volumes at around 2.5 Tb (trillion base pairs, so about 5 TB) during a week-long run across the 12 Illumina instruments in operation today. Additionally the JGI has two Pacific Biosciences instruments producing data exponentially (5X) faster every year, and presently each create 20 TB of data/year. The total data output (raw and processed) from these platforms exceeds 5 PB/year transferred to the National Energy Research Scientific Computing Center (NERSC). After the raw image data are processed at NERSC, a reduced set of raw sequence data is transmitted to the National Center for Biotechnology Information's (NCBI's) Sequence Read Archive (SRA). The raw sequence data are further analyzed, assembled and/or processed at NERSC by project-specific computational pipelines, and delivered to the collaborators for each project. JGI has also recently increased its computing capacity through the help of ARRA funding and more recent purchases. Clusters to support assembly and analysis are presently at roughly 4,000 cores and will double in the next few weeks as new hardware comes online. In addition, around 20 specialized large-memory nodes support the most complex and demanding analysis workloads. The JGI has almost 3 PB of data online in the file systems. Critical data is backed up in NERSC's High Performance Storage System (HPSS) as a redundant, offline copy. Additionally, long-term experimental raw and results data sets are permanently archived at HPSS to free up filesystem resources.

Since late 2010, the IT infrastructure, clusters, filesystems, and networking for JGI has been primarily located at NERSC (the Oakland Scientific Facility) and under its management. The production sequencing continues to be performed at the Walnut Creek Facility and the data are transferred over ESnet to NERSC for storage and analysis over a portion of a dedicated 10 Gbps link. Limited computing resources remain at

Walnut Creek to ensure sequencing runs smoothly and to provide temporary caching of data. Additionally, desktops and workstations remain at Walnut Creek, some of which require interactive analysis and visualization of the data sets stored at NERSC.

JGI is slated for relocation to Richmond as one of the first buildings in LBNL's second site in 2017. It is expected that the instruments that produce the raw data will remain at the JGI in Richmond and that the clusters and file systems that analyze and store the data will remain at NERSC, which is planned to be at the LBNL primary site in Berkeley by that time.

16.2.2 Process of Science

JGI supports and performs genomic science by sequencing, synthesis, and data analysis, as well as utilizing the end-product data for scientific insight. JGI accepts Community Sequencing Proposals for plant, microbial, fungal, and other organisms. These sequencing proposals include single-organism, single-cell, and environmental or metagenome sequencing. Accepted proposals ship wet samples to JGI for sequencing, assembly, and analysis. Users access and analyze the data for completed sequences through data portals such as the Genome Portal. In addition, the data are typically uploaded to the NCBI SRA for broader access. Data can also be analyzed through various pipelines such as IMG, Phytozome, and the Systems Biology Knowledgebase (KBase). The results of this analysis are made available through direct transfer, Globus Online, or through data portals specific to the pipelines. In addition to analyzing sequence data produced at JGI, external users can submit sequence data to these pipelines for analysis and publishing. Finally, these data products are used by research groups at JGI for scientific insight.

16.3 Key Remote Science Drivers

16.3.1 Instruments and Facilities

Some key JGI users are the BRCs at LBNL, ORNL, and University of Wisconsin-Madison; however, they only amount to less than 10% of the projects. A wide variety of universities, government agencies, and research institutes also utilize JGI sequencing resources. Specifically, we anticipate that collaborations with the University of California at Davis will accelerate in the coming years as one of our Big Data projects, resequencing several thousand strains of crops, and directly involving collaborators located there.

16.3.2 Process of Science

JGI project collaborators ship wet samples to JGI for sequencing and analysis. The data are then accessed through portals such as the Genome Portal. Depending on the type of sample and the preferences of the customer, the data may also be uploaded to NCBI.

In addition to shipping wet samples for sequencing, remote users also submit sequenced data to JGI for analysis by the various pipelines, typically through a Web

portal. The resulting analysis is either viewed through the portal or potentially downloaded. Presently, these data sets are typically 1 GB–20 GB. However, there may be hundreds to thousands of submissions. Since these are submitted via the Web, the connection may vary from 100 Mbps to 10 Gbps at the remote site. Periodically, data-set exchanges can be much larger than this (i.e., 100 GB–3 TB) with commercial vendors or international collaborators and they almost always require shipment of a multi-TB USB hard drive, as the remote site lacks the bandwidth to transfer that amount of data reliably or efficiently.

There are other risks and external influences of JGI's network requirements of which the impact is not well known. For example in 2011, NCBI announced (and then retracted) that it would shut down the SRA and possibly other services. Because the JGI relies on NCBI to comply with its mandate to publicly release its data, it would have to replace that service with its own implementation. This would significantly impact network usage patterns, as these public data sets are repeatedly searched and downloaded across the world. Additionally, KBase is proposing to create an assembly service that could easily result in an additional 1 TB of data per project migrating through JGI networks.

16.4 Local Science Drivers — the Next 2-5 Years

16.4.1 Instruments and Facilities

Second-generation sequencing platforms from Illumina, and third-generation platforms from Pacific Biosciences and, potentially, Oxford Nanopore are expected to increase sequencing rates while reducing the cost to sequence. In the past few years, sequencing costs have dropped by over an order of magnitude as measured on a $/base cost, far eclipsing Moore's law. This 5X per year trend is expected to continue, especially for the third-generation sequencers, and will likely reduce to just 2X per year for the second-generation platforms. Consequently, facilities such as JGI are grappling with the expected growth for data generated locally as well as the volume of data submitted from external users. JGI expects to continue to maintain and update its sequencers in the coming years and thereby benefit from further $/base reductions. This will significantly increase the data volumes that must be assembled, analyzed, and eventually made public.

JGI, in collaboration with the NERSC Division, will operate computing systems to support this analysis. However, as the ingest rate increases, data rates may outgrow JGI's computing resources. Therefore, JGI is also evaluating how it can leverage HPC resources at NERSC and the Leadership Computing Facilities (LCFs) (ANL and ORNL) to help meet this need. Furthermore, members of JGI in collaboration with the KBase program are exploring how cloud computing may be used to augment local resources. If cloud computing becomes a critical component of the analysis infrastructure for JGI, network patterns could change significantly.

16.4.2 Process of Science

The overall process of sequencing, assembly, analysis, and annotation will largely remain the same. JGI plans to perform more synthetic, metagenomic, and functional science in the coming years, which will continue to drive the need for more sequencing.

16.5 Remote Science Drivers — the Next 2-5 Years

16.5.1 Instruments and Facilities

It is anticipated that the lower sequencing cost will lead to a rapid growth in remote sites operating sequencers and submitting data to JGI for analysis. This will translate into more data being transferred across the wide-area network. As computational demands grow, sites may start to utilize cloud computing. This would impact how data flows among remote sites, collaborators, JGI, and cloud providers.

16.5.2 Process of Science

It is not anticipated that the overall process will change.

16.6 Beyond 5 Years — Future Needs and Scientific Direction

If current trends continue in sequencing technology, there could be increasing demands on the network. It is expected that more sequence data will come from external facilities and, consequently, JGI will be required to import and process more data.

16.7 Network and Data Architecture

JGI and KBase have been using Globus Online to transfer large data sets between Argonne and NERSC with great success.

16.8 Collaboration Tools

As travel budgets have tightened recently, we anticipate an increased use of ReadyTalk, Google Talk, and/or Skype. The JGI uses ReadyTalk almost daily during regular meetings with remote personnel. Since the JGI maintains hundreds of collaborators across North America, South America, Europe, Asia, and Australia, we have increasingly used Google Talk and Skype to facilitate communications.

16.9 Data, Workflow, Middleware Tools, and Services

JGI currently relies primarily on in-house developed portals and pipelines to manage workflows. Some of these do use third-party developed tools (i.e., Genome Browser, SGE, etc.) but most of these do not have a significant network component. Data transfers are typically performed automatically between local facilities or interactively through a Web interface or FTP site.

16.10 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• Illumina • Pacific Biosciences	• Data sequenced and processed at JGI/NERSC • Genome annotation at NERSC • Uploaded to NCBI & remote collaborators	• JGI to NERSC • 7-8 TB/day • 1.5 million files	• Avg: 7-8 TB/day • Peak: 10 TB/day • Multiple 1 Gbps and 10 Gbps connections	• Average transfer of 3 TB/week • Expected bandwidth requirement 1-10 Gbps (in bursts) •
• Portals • KBase • Collaborators	• Interactive Web and analysis		• Multiple 1 Gbps links to filesystems and databases	• Greater than 1 Gbps bandwidth needed • High availability needed
• NCBI	• Data repository			• 1 Gbps to 10 Gbps of bandwidth
• UC Davis	• Big Data projects collaborator			• 1 Gbps or higher bandwidth
• Communication and collaboration tools (GTalk, Skype, etc.)	• Increasingly important as travel budgets are reduced			
2-5 years				
• Illumina • Pacific Biosciences • Oxford	• 50X to 250X growth of data from 2012 • Data sequenced and processed at JGI/NERSC • Genome annotation at NERSC • Uploaded to NCBI and to remote collaborators	• 750 TB/day • 150 million files	• Average transfer of 700 TB/day • Peak transfer of 1 PB/day	• Average transfer of 300 TB/week
5+ years				
• Illumina • Pacific Biosciences • New 4th generation sequencers	• 1,000x to 5,000x growth of data from 2012 • Data sequenced and processed at JGI/NERSC • Genome annotation at NERSC • Uploaded to NCBI and to remote collaborators	• 7 PB/day • 1 billion files	• Average 7 PB/day • Peak 10 PB/day	• Average transfer of 3 PB/week

Note: It is really hard to project five years in this genomic sequencing industry. At this scale, we will have to adjust the local network topology and re-engineer workflows.

17 KBase — the Systems Biology Knowledgebase

17.1 Background

The Systems Biology Knowledgebase (KBase) is a multi-institutional effort to build an integrated knowledgebase and data analysis facility for data reflecting biological systems. The system will be instantiated as a service-oriented architecture spanning multiple systems distributed across sites; each system has a discrete focus for system infrastructure and analytical processes. Our plan is to use ESnet for two major purposes: (1) ESnet will provide connectivity to users, both within DOE as well as on to the Internet at large. (2) ESnet will function as a high-speed backplane, connecting system components distributed across multiple sites. In the first case, raw data and data products will be transferred between the various systems in KBase for analysis and fault tolerance. In the latter case, users will submit raw data for analysis and integration into KBase.

The core KBase architecture is a set of services that provide access to high-value data products as well as analytical services. Discrete analysis and modeling services (referred to as “cores”) are deployed in a redundant fashion across KBase sites. KBase service calls can result in heavyweight analysis/modeling application execution. These applications each run on one or more of the sites, with data and results moved using Globus Online.

17.2 Key Local Science Drivers

17.2.1 Instruments and Facilities

KBase uses several systems distributed across the various sites. At Argonne, KBase uses the Magellan system (7,000 cores, 20 TB memory, demonstrated capability to use 100 GE effectively). At LBNL/National Energy Research Scientific Computing Center (NERSC), KBase has an allocation on Carver. The Kandinsky system at ORNL will be used for data-intensive workloads. Finally, a small-scale system at BNL is dedicated to KBase. Some workloads will be run on Mira at Argonne and Franklin at NERSC. KBase is partnering with the Joint Genome Institute (JGI), which will result in large-scale ingestion of JGI-produced sequence data to provide novel KBase analysis and modeling of these data sets.

17.2.2 Process of Science

The KBase project is building novel analysis and modeling techniques for biological data, focusing on microbes, plants, and microbial communities, as well as a service-oriented architecture that delivers analysis and modeling services to users. Users either upload their own data sets or make use of data sets already loaded into KBase, and apply KBase operations to these data sets.

Developers can also develop new analysis and modeling approaches and integrate them into KBase. The goal here is to provide a common infrastructure for large-scale

biological data analysis and model creation and refinement. Improvements developed through these processes will be rolled out for KBase users over time.

17.3 Key Remote Science Drivers

17.3.1 Instruments and Facilities

The distributed architecture of KBase will be a major driver of its consumption of network services. As the aggregate volume of data handled by KBase grows, its use of ESnet for wide area transfers will grow proportionately.

Similarly, as the user base of KBase grows, its data ingestion footprint will grow. The beta release of KBase will occur in February 2013; it will be difficult to gauge the growth of the KBase user community until after that release. Initial users are from the various BRCs, other DOE facilities, and BER grantees.

17.3.2 Process of Science

Wide area data transfers will largely provide computational infrastructure; we don't have plans to do remote visualization or network-intensive collaboration activities at the moment.

17.4 Local Science Drivers — the Next 2-5 Years

17.4.1 Instruments and Facilities

We expect that hardware refreshes will occur at the KBase sites throughout the course of the project.

17.4.2 Process of Science

The scalability, sophistication, and number of methods available in KBase will increase greatly over the next few years. To keep pace with growth in data-set sizes (described below), techniques will need to be scaled to data sets multiple orders of magnitude larger than today. Each new technique will interest potentially different user communities, depending on the particular technique. For example, our variation analysis pipeline is of great interest to researchers studying plant genomes; it is being used to study a large resequencing study of poplar genomes. This approach wouldn't be particularly applicable to researchers studying microbial communities.

It is difficult to predict precisely which new methods will become popular with users at this stage. However, broadly speaking, we can describe our techniques in terms of two classes. Some approaches consume relatively small amounts of data, for example fully assembled microbial genomes. These data sets are on the order of tens of MB of data. Other approaches require the ingestion of full raw sequencing data sets. These data sets are growing quickly. The class of applications that use full raw data sets, as opposed to smaller data products, will largely define our wide area network footprint.

17.5 Remote Science Drivers — the Next 2-5 Years

17.5.1 Instruments and Facilities

As KBase moves into production, the demands it places on the distributed backplane connecting core sites with infrastructure will increase. This will largely fill the role of data-movement infrastructure.

We expect that the current trend of sequencer democratization will continue, which will produce increasingly large quantities of data sets from a diverse set of remote sites.

17.5.2 Process of Science

We expect the core architecture (applications run at single sites, federated by wide area data movement, with centralized integrations into core KBase services) to continue in the same basic form. New methods requiring more data or producing larger results will clearly result in increased demands on the network.

17.6 Beyond 5 Years — Future Needs and Scientific Direction

It is unclear what network requirements KBase will have beyond the five-year horizon.

17.7 Network and Data Architecture

We will be using Globus Online for wide area data transfers. This will be deployed in a fairly vanilla fashion; we will be using purpose-built DTNs, either in discrete hardware configuration or in a virtualized environment on Magellan. The network of DTNs will provide intersite movement of data sets for analysis and synthesis, as well as integration using KBase models.

17.8 Data, Workflow, Middleware Tools, and Services

KBase can be thought of as a project to build a service-oriented architecture around analysis and modeling workflows. A key difficulty is the rapid growth of sequence data. Costs for such data sets have been dropping at a rate approximating 10X a year for the past five years, a rate far faster than Moore's law. This in turn has caused two related events to occur. First, the accessibility of sequencer ownership has greatly increased, democratizing the process of data-set creation. We expect this fact to result in data sets from many facilities with no previous local sequencing capability. Second, the reductions in cost are greatly increasing the volume of data being produced, both in aggregate quantity as well as numbers of discrete samples.

In many ways, the KBase architecture is a direct reaction to biology becoming a data-rich science after decades of dependence on low-throughput, effort-intensive data-collection methods. To improve the state of the art in this area, we need to encode best analysis practices in the form of workflows, as well as make data aggressively reusable to keep computational costs under control. This approach is highly dependent on high-

performance networks (and specifically ESnet), as KBase is only exists virtually; all users are remote.

We don't have plans to use commercial cloud offerings at the moment.

17.9 Outstanding Issues

As we mentioned earlier, KBase cores that consume full raw data sets will dominate the network footprint of the project. While it is difficult to accurately predict the eventual demand for each of these cores, we can use a popular pre-existing KBase core to estimate. MG-RAST (Metagenomics Rapid Annotation using Subsystem Technology) is the primary metagenomic (microbial community) annotation portal. This project has been in operation since 2007 and has experienced substantial growth over the past few years. Over the past two months, MG-RAST annotated 5,500 date sets, totaling 2 terabasepairs of raw data, more like 3 TB of full data including quality scores and associated data. Between upload of these data sets, download of results, and remote analysis of results via the system APIs, this system probably consumes a total of 5-6 TB of network bandwidth per month. We are seeing year-over-year doubling of data analysis volumes.

KBase will include several cores that directly process full-sequence data sets. Today, there are metagenomics annotation services, variation analysis, genomic and metagenomic assembly, and quality-assessment services. Each of these cores will likely have similar bandwidth requirements and growth patterns, as they grow to similar maturity as the annotation service. More of these cores will be developed over time.

Openflow is of interest. We are investigating local deployment for Magellan at Argonne.

17.10 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• Systems at the four KBase sites • Data from JGI and other large- and small-scale sequencing centers	• User submission of data • Data analysis and model construction process distributed across sites • Results centrally integrated	• 10-100 TB/day • Full sequence data sets 18 files of 35 GB (max) • 20-200 data sets a day	• No turnaround time requirements • Most uses async	• 70% of data will be moved between ANL, LBNL, and ORNL
2-5 years				
• Growth of user community	• New analysis/model construction methods added	• 100 TB-2 PB/day • Full-sequence data sets 18 files of 350 GB (max) • 50-300 data sets a day		• Increasing data to end user sites
5+ years				
• Growth of user community	• New analysis/model construction methods	• 2 PB+? • Continued growth of sequence data sets		

18 Microbial Genome Sequencing Projects: Environmental and Population Studies

18.1 Background

Next-generation sequencing has enabled researchers to perform genomic and transcriptomic sequencing at rates unimaginable in the past. Microbial genomes now can be sequenced in a matter of hours, leading to a significant increase in the number of assembled genomes being deposited in the public archives. This huge increase in DNA sequence data presents new challenges for the submission, annotation, and analysis pipelines. New standards for the submission, validation, and analysis of genome data must be developed for both reference genomes and environmental and population studies derived from clinical outbreaks.

New data formats need to be developed to address high volume and high redundancy of the data for data storage, retrieval, and exchange. The current National Center for Biotechnology Information (NCBI) collection of proteins contains more than 20 million sequences.

The analysis of the microbial community requires a complex system approach capturing metadata such as information about the biological sample, habitat (ecological and medical data), biotic relationship, and more. Collecting and linking metadata with genomic (and other omics) data is needed.

18.2 Key Local Science Drivers

18.2.1 Instruments and Facilities

Genome and metadata at NCBI; assembled microbial genomes in GenBank — 7,500; in SRA — 55,000

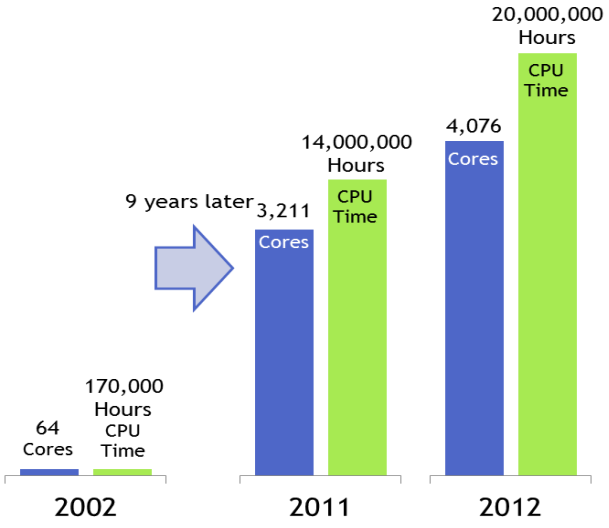


Figure 25. Compute growth

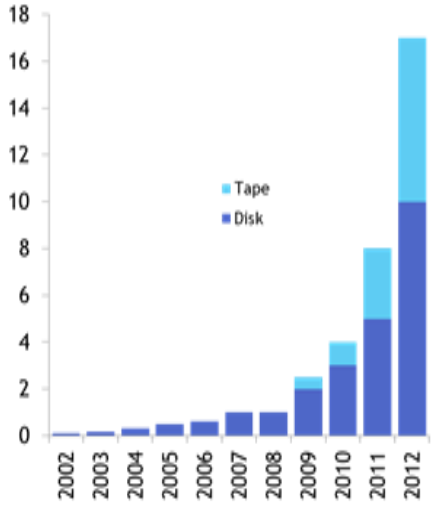


Figure 26. Storage growth in petabytes

18.2.2 Process of Science

- Primary submission and dissemination of raw data: primary archive
- Reduced redundancy, pre-computed results of analysis: Refseq
- BioProject, BioSample, Sequence Read Archive (SRA), genome Assembly; metagenome projects; annotation.
- Validation (contamination screen, 16S, phylogenetic analysis)
- Compression in SRA – see Big data initiative
- Reduce redundancy: pan-genome and protein clusters

18.3 Key Remote Science Drivers

18.3.1 Instruments and Facilities

- Submission Portal
- Entrez search and retrieval system
- FTP access to sequence data

18.3.2 Process of Science

- Submission, validation, storage, distribution – public archives
- Analysis, protein clusters, pan-genome, annotation – Refseq project

18.4 Local Science Drivers – the next 2-5 years

18.4.1 Instruments and Facilities

Genome sequencing is rapidly advancing. Raw sequence data will be replaced with more usable and storable formats produced directly by sequencing instruments.

Tools will include:

- fast comparison
- reduced redundancy
- fast retrieval
- new data transfer protocols

18.4.2 Process of Science

With the rapid growth of genome sequence data the data model will probably change. Only reference genomes will be assembled and submitted to GenBank.

The differences (variations) will be stored as alignments to the reference (BAM files). Huge redundancy in protein data set has to be addressed. Anew data model is needed – pan-genome may be a solution. NCBI is proposing a super-gi model for the nearest ~ 2 years future. Identical proteins within the same species will re-use the same gi number.

18.5 Remote Science Drivers — the Next 2-5 Years

18.5.1 Instruments and Facilities

SRA (Sequence Read Archive) will be modified. There is no need to capture and store all the reads coming out from the sequencing machines. Public archives should store the results of analysis. The model is yet to be defined.

Example

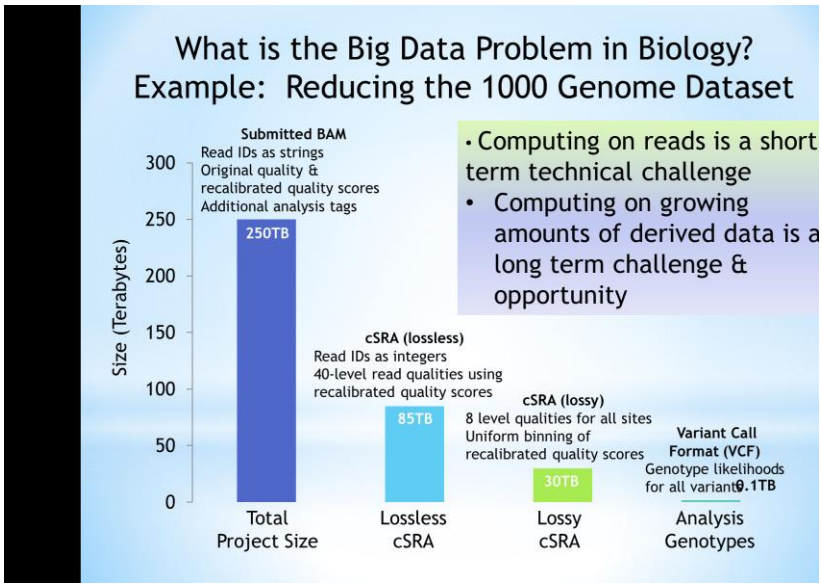


Figure 27. Reducing the 1000 Genome Data Set

18.5.2 Process of Science

Submission, validation, storage, distribution – public archives

Analysis, protein clusters, pan-genome, annotation – Refseq project

Several changes are planned in the process of science – proposed by NCBI awaiting INSDC approval:

- Taxonomy: taxid will no longer be assigned below species
- Level: sequence data will be identified by BioSample ID, BioProject ID, SRA experiment; genome assembly is uniquely identified by Assembly ID

18.6 Beyond 5 Years — Future Needs and Scientific Direction

Genome data storage and comparative analysis tools cannot keep up with the rate of genome sequencing. Probably central resource of precomputed results with easy access is needed. Genome sequence data should be integrated with habitat (ecology, geochemistry); phenotype; metabolic pathways. Interactions of all members of an ecological niche (bacteria, viruses, eukaryotes) should be captured to understand the life of the community, not individuals.

18.7 Collaboration tools

With the restrictions on travel, remote collaboration tools become increasingly important.

18.8 Data, Workflow, Middleware Tools, and Services

Raw sequence reads (SRA) if still needed can be moved to the cloud. Search and retrieval protocols should be modified (FTP is too slow, consider Aspera - http://asperasoft.com/)

18.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• Genome sequence data in SRA and GenBank	• Submission, validation, distribution • Data compression	• Data volume • FTP downloads: 26.6 TB/day		
2-5 years				
• Metadata, pre-computed results of analysis	• Submission, validation, distribution • Reduce redundancy by analyzing the data and distribute the results, not raw data			
5+ years				
•	• Replace genome sequences and raw data			

19 Pan-omics Facility, PNNL

19.1 Background

Proteomics is the process by which the complement of proteins expressed in a cell or a subset of this complement is identified, quantified, and utilized for biological discovery. In the same vein, metabolomics and lipidomics is the process by which small metabolites and lipids are utilized in the same way with the ultimate purpose of integrating the disparate data types into one biological model. For proteomics, the proteins are cleaved into smaller pieces called peptides and analyzed in a mass spectrometer that converts the biological polymer consisting of amino acids to a distinct mass and fragmentation pattern that can be represented numerically. These values are then compared back to a sequenced genome, metagenome, or collection of transcript sequences to determine the amino acid sequence. Similarly, the mass spectral output for metabolites and lipids, which consist of masses and elution times, are compared back with known metabolite databases for identification. Once identified, the protein (as correlated through the identified peptide), metabolite, and lipid and the correlated biological function and pathway are integrated into either regulatory or metabolic models for biological elucidation of the organism or community.

19.2 Key Local Science Drivers

19.2.1 Instruments and Facilities

Our facility consists of more than 30 high-throughput measurement platforms that have the capacity to run 24 hours a day/7 days a week.

For data processing and analysis, we have more than 250 computers consisting of 600 computational cores that are connected through high-speed network connections internal to the PNNL system. These systems are PC-class machines that are refreshed every three to four years.

We utilize about 80 TB of working storage space with over 300 TB of data in an archive. We are generating about 3 TB/month of data on the systems at the moment.

We have a connection to the Environmemtal Molecular Sciences Laboratory (EMSL) supercomputer and a campus-wide high-speed Internet connection.

We use PRISM (Pan-omics Research Information Storage and Management), a data-management system that collects data files from multiple mass spectrometers and manages the storage and tracking of these data files, automating their processing into intermediate results and final products. It collects and maintains information about the biological samples used in research experiments and the laboratory protocols and procedures used to prepare them. The system also allows users to locate and examine the data that it contains, and allows other information systems to access appropriate portions of it. The final products of the system are compilations of peptides observed in the biological samples for a particular organism under specified sets of conditions

chosen by the researchers in their experiments. Researchers query these compilations to determine what proteins an organism creates under different conditions of growth and stress in order to better understand the how cellular biological mechanisms work.

19.2.2 Process of Science

As stated above, through measurements in the mass spectrometer, the information on the biological samples is converted to numerical values, which relate back to either a particular biological peptide sequence or a metabolite. By relating these values back to a database (either a genome sequence or a metabolite library), the mass spectral outputs are identified where the peak intensities measured by the mass spectrometer relate to the abundance of the biomolecule in the sample. Once the identifications and quantities are known, the peptides and metabolite abundances can be compared between samples and inserted into metabolic or regulatory models or reported back to the biologists directly. At this point, human intervention must relate the information into biological insight or discovery.

19.3 Key Remote Science Drivers

19.3.1 Instruments and Facilities

See *Instruments and Facilities* above.

19.3.2 Process of Science

Outside of PNNL, we utilize both genome sequences and metabolite libraries that can be either downloaded from Web sites or sent via e-mail. Once mass spectral data have been generated at PNNL, the data are stored in our archive indefinitely and processed into either peptide or metabolite lists and abundances. This reduction of data is necessary for a number of reasons:

1. For biological interpretation, biomolecule identification is always the necessary first step, and it is easier for the collaborator (usually a biologist) to understand the list of biomolecules rather than the raw mass spectra.
2. The size of the raw data is usually too large to transfer directly. The biggest issue here is the bandwidth needed to transfer the data, usually on the collaborator side. We find that filling a hard drive with raw data and shipping it directly is far more effective than direct data transfers.
3. There is no universal translator for data. This means that unless you have the software that can analyze our particular types of raw data, it isn't much use, whereas the lists of biomolecules are usually in tab- or comma-delimited form, or in an Excel format that can be used almost universally. Better interoperability is needed between tools.

19.4 Local Science Drivers — the Next 2-5 Years

19.4.1 Instruments and Facilities

We expect the number of instruments to increase nominally (maybe by 20%) but the throughput to increase, thereby increasing our data generation. We expect data generation to increase by 10X.

Our computational platforms are refreshed every three to four years, and we increase the number of these as the need arises. As the performance of new machines and the efficiency of the processing software algorithms increase, the need for more machines decreases. This is offset by an increase in the throughput, data density per data set, and types of data sets collected, increasing the need for computational power. Predicting this for the next two to five years is difficult, but best guess is 2X in about five years.

19.4.2 Process of Science

The basic premise for biological discovery will not change dramatically over the next two to five years. The main changes will be in the complexity of the samples analyzed, which increases the need for computational power, and the next generation of instruments, which will be faster, more sensitive, and with a higher dynamic range, thus increasing the data density per data set.

19.5 Remote Science Drivers — the Next 2-5 Years

19.5.1 Instruments and Facilities

See *Instrumentation and Facilities*, above.

19.5.2 Process of Science

As we progress with the Pan-omics paradigm, the ability to integrate the data and work in a collaborative mode will become more important. The largest impediments to the collaborative data-sharing model are the firewalls at the different institutions and the interoperability of tools to analyze the data. Tools like DropBox are nice, but certain institutions (like PNNL) have blocked their use as a collaborative tool. What is needed is a DropBox or cloud-type infrastructure that is extremely user friendly and can be shared by multiple institutions. These tools should be able to store different types of data at different levels of processing while having the safeguards needed to ensure privacy and security. What is ultimately needed is a universal cloud or “C” drive, which everyone can access, that stores massive amounts of processed and unprocessed data.

19.6 Beyond 5 Years — Future Needs and Scientific Direction

The largest change to the Pan-omics pipeline would be the implementation of our Ion Mobility Mass Spectrometers on a massive scale that would allow 100 samples to be analyzed in parallel. This would increase our data streams by orders of magnitude and would come with the commensurate data computing needs.

19.7 Network and Data Architecture

PNNL currently has a data transfer node (dtn.pnl.gov) attached to its Secure Collaboration Zone (SCZ) perimeter network. The dtn.pnl.gov system has a 10Gbps connection to the Internet and QDR IB attached to a 4 PB Lustre storage cloud; it supports 1 GB/sec data transfers. The storage cloud has multiple internal mount points, and is available to the Olympus supercomputer via QDR IB interconnects. EMSL also has its Aurora archive attached to the SCZ network at 10Gbps, providing up to 1 GB/sec data transfer capability to other laboratories. The SCZ has a perfSONAR/NDT testing point attached at 10 Gbps (ndt.pnnl.gov). The SCZ utilizes a host-based firewall model in which a Port Scan Attack Detector (PSAD) is used in conjunction with iptables to detect and block attackers with little performance degradation on individual hosts.

For collaborative research, we envision an environment where data transfers, when needed, occur with the ease of sending an e-mail, with rapid data uploads and downloads on both sides of the connection. While this will require many institutions to change their architecture, a more reasonable approach would be to have a cloud-like access paradigm to the data, where data at all levels and tools to interact with the data co-exist on the cloud and collaborators can interact with the data without needing to download it. This would also require the interoperability of data formats and analysis tools.

19.8 Collaboration Tools

Our institution has made it very difficult to use the collaboration tools that have become commonplace in the industry and beyond. We are limited to using Skype from computers that are personal and not on the internal network, unless we are off site, where we can use a company computer on an external server. DropBox is blocked, as are other collaborative tools. The largest impediment to collaboration is not computational or technological. It is policy. We teleconference by phone or use programs like ReadyTalk and Lync for group presentations.

19.9 Data, Workflow, Middleware Tools, and Services

Our internal data management will scale with the increased data production and data growth. The largest need will be in the translation of the information into biological insight and the ability to integrate the data into models that can be interpreted by humans.

19.10 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• Mass-spectrometry-based Pan-omics measurements. Mostly from commercial-grade instruments.	• Mass spectrometry results are compared to a database for identification and quantitation. These results are then integrated into biological models.	• 3 TB/month • Each data set is 0.5 GB to 10 GB	• 10 GB in a couple of minutes	• Usually done through e-mail or sending hard drive
2-5 years				
• Ion Mobility Mass spectrometry with advanced separations	• No change	• 30 TB/month • Each data set is 10 GB to 30 GB	• 50+ GB in a couple of minutes	• Usually done through e-mail or sending hard drive
5+ years				
• Massively parallel mass spectrometry	• No change	• 3 TB/day • Each data set is 10 GB to 30 GB	• 100+ GB in a couple of minutes	• Hopefully done through a common cloud

20 Glossary

GB/sec: Gigabytes per second – a measure of network bandwidth or data throughput

Gbps: Gigabits per second – a measure of network bandwidth or data throughput

MB/sec: Megabytes per second – a measure of network bandwidth or data throughput

Mbps: Megabits per second – a measure of network bandwidth or data throughput

PB/sec: Petabytes per second – a measure of network bandwidth or data throughput

Pbps: Petabits per second – a measure of network bandwidth or data throughput

TB/sec: Terabytes per second – a measure of network bandwidth or data throughput

Tbps: Terabits per second – a measure of network bandwidth or data throughput

AIRS	Atmospheric Infrared Sounder
ALCF	Argonne Leadership Computing Facility
AMIP	Atmospheric Model Intercomparison Project
ANA4MIPS	Reanalysis for the Coupled Model Intercomparison
ANL	Argonne National Laboratory
API	application programming interface
APS	Advanced Photon Source
AR4	Fourth Assessment Report
ARM	Atmospheric Radiation Measurement
ARMBE	Atmospheric Radiation Measurement Best Estimate
ARRA	American Recovery and Reinvestment Act
ASCR	Advanced Scientific Computing Research
BADC	British Atmospheric Data Centre

BER	Biological and Environmental Research
BES	Basic Energy Sciences
BG/Q	Blue Gene/Q
BNL	Brookhaven National Laboratory
BRC	Bioenergy Research Center
CA	cooperative agreement
CAM	Community Atmosphere Model
CCSM3	Community Climate System Model, Version 3
CCSP	Climate Change Science Program
CDIAC	Carbon Dioxide Information Analysis Center
CEDA	Comprehensive Environmental Data Archive
CEMS	Climate and Environmental Monitoring from Space
CERES	Clouds and the Earth's Radiant Energy System
CESD	Climate and Environmental Sciences Division
CESM	Community Earth System Model
CET	Center for Enabling Technologies
CF	Climate Forecast
C-LAMP	Carbon Land Model Intercomparison Project
CM	Climate Model
CMIP	Coupled Model Intercomparison Project
CORDEX	Coordinated Regional Climate Downscaling Experiment
CSP	Community Sequencing Program
CSS	Climate Storage System
CSSEF	Climate Science for a Sustainable Energy Future
DCMIP	Dynamical Core Model Intercomparison Project
DDC	Data Distribution Center
DKRZ	German Climate Computational Center
DMF	Data Management Facility
DNS	Domain Name System
DOE	Department of Energy
DOI	digital object identifier

DTN	data transfer node
ECMWF	The European Centre for Medium-Range Weather Forecasts
EDEN	Exploratory Data analysis Environment
EMSL	Environmental Molecular Sciences Laboratory
ENES	European Network for Earth System Modeling
EO	Earth observation
EPIC	Environmental Policy Integrated Climate
ESGF	Earth System Grid Federation
ESM	Earth System Modeling
ESnet	Energy Sciences Network
EUCLYPSE	European Union Cloud Intercomparison, Process Study & Evaluation Project
FRE	FMS Runtime Environment
FT	Fourier transform
FTICR	FT ion cyclotron resonance
FTP	File Transfer Protocol
GCAM	Global Change Assessment Model
GCE	Green Collaboration Environment
GCM	global circulation model
GCRA	Global Change Assessment Model
GDO	Green Data Oasis
GeoMIP	Geo-engineering MIP
GFDL	Geophysical Fluid Dynamics Laboratory
GIS	geographic information system
GLBRC	Great Lakes Bioenergy Research Center
GPU	Graphics processing unit
GUI	graphical user interface
HDF	Hierarchical Data Format
HECToR	High-End Computing Terascale Resource
HPC	High-performance computing
HPLC	HPSS High Performance Storage System

HPSS	High Performance Storage System
HRMAC	High Resolution and Mass Accuracy Capability
HTTP	Hypertext Transfer Protocol
IA	integrated assessment
IB	InfiniBand
ICE	initial condition ensemble
IMG	Integrated Microbial Genomes
IMS	ion mobility spectrometry
I/O	input/output
IPCC	Intergovernmental Panel on Climate Change
ISME	International Society for Microbial Ecology
ISP	Internet service provider
JASMIN	Joint Analysis system Meeting Infrastructure Needs
JAXA	Japan Aerospace Exploration Agency
JGCRI	Joint Global Change Research Institute
JGI	Joint Genome Institute
KBase	Systems Biology Knowledgebase
KNMI	Royal Netherlands Meteorological Institute
LAN	Local area network
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LC	Livermore Computing
LCA	life-cycle assessment
LCF	Leadership Computing Facility
LDAP	Lightweight Directory Access Protocol
LIGO	Laser Interferometer Gravitational Wave Observatory
LLNL	Lawrence Livermore National Laboratory
LTQ	linear trap quadrupole
LUCID	Land-Use and Climate, Identification of Robust Impacts
MERRA	Modern Era Retrospective Analysis for Research and Applications
MG-RAST	Metagenomics Rapid Annotation using Subsystem Technology

MIP	model intercomparison project
MISR	Multi-angle Imaging SpectroRadiometer
MLS	Microwave Limb Sounder
MPI	Message Passing Interface
MONSooN	Met Office and NERC Supercomputing Node
NARCCAP	North American Regional Climate Change Assessment Program
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCAS	National Centre for Atmospheric Science
NCBI	National Center for Biotechnology Information
NCEO	National Centre for Earth Observation
NCL	NCAR Command Language
NERC	Natural Environment Research Council
NEODC	NERC Earth Observation Data Centre
NERSC	National Energy Research Scientific Computing Center
NetCDF	Network Common Data Form
NDT	Network Diagnostic Tool
NFS	network file system
NGE	Next Generation Ecosystem
NIH	National Institutes of Health
NMR	nuclear magnetic resonance
NOAA	National Oceanic and Atmospheric Administration
NRC	Nuclear Regulatory Commission
NSA	North Slope of Alaska
NSF	National Science Foundation
NWSC	NCAR-Wyoming Supercomputing Center
obs4MIPs	Observational Products More Accessible for Coupled Model Intercomparison Projects
OLCF	Oak Ridge Leadership Computing Facility
OpenDAP	Open-source Project for a Network Data Access Protocol
ORNL	Oak Ridge National Laboratory

OSCARS	On-Demand Secure Circuits and Advance Reservation System
PCMDI	Program for Climate Model Diagnosis and Intercomparison
pdf	probability density function
perfSONAR	PERformance Service Oriented Network monitoring Architecture
PF	petaflop
PMIP3	Paleoclimate MIP Phase 3
PNNL	Pacific Northwest National Laboratory
POP	Parallel Ocean Program
POTS	plain old telephone service
PPE	perturbed physics ensembles
PRISM	Pan-omics Research Information Storage and Management
PSAD	Port Scan Attack Detector
QC	quality control
QDR	quad data rate
RAL	Rutherford Appleton Laboratory
RDHPCS	Research and Development High Performance Computing System
REST	Representational State Transfer
RGCAM	Regional Global Climate Assessment Model
SaaS	Software as a Service
SAN	Storage area network
SC	DOE Office of Science
SciDAC	Scientific Discovery through Advanced Computing
SCP	secure copy
SCZ	Secure Collaboration Zone
SGP	Southern Great Plains
SRA	Sequence Read Archive
SSH	Secure Shell
STFC	Science and Technology Facilities Council
TAMIP	Transpose-AMIP
TCP	Transmission Control Protocol
TIC	Trusted Internet Connection

TOF	time of flight
TRMM	Tropical Rainfall Measuring Mission
TVN	Thames Valley Network
UC	University of California
UCAR	University Corporation for Atmospheric Research
UI	user interface
UK	United Kingdom
UQ	uncertainty quantification
UTEM	ultrafast transmission electron microscopy
UV-CDAT	Ultrascale Visualization-Climate Data Analysis Tools
VAP	value-added product
VOIP	voice over Internet Protocol
VNC	Virtual Network Computing
VPN	virtual private network
WAN	wide area network
WCRP	World Climate Research Program
WDDC	World Data Center for Climate
WGCM	Working Group on Climate Modeling
WMO	World Meterological Organization
XDC	External Data Center

21 Acknowledgements

This work would not have been possible without the contributions and participation of those who provided information and attended the review. ESnet would also like to thank the BER program office for their help in organizing the review and providing insight into the facilities supported by the BER program. In addition, the Oak Ridge Institute for Science and Education (ORISE) conference support and logistics staff was very helpful.

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR). Vince Dattoria is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract DE-AC02-05CH11231.

This work was supported by the Directors of the Office of Science, Office of Advanced Scientific Computing Research, Facilities Division, and the Office of Biological and Environmental Research.

This is LBNL report LBNL-6395E