



U.S. Department of Energy's
Office of Science

Advanced Scientific Computing Research Program

Distributed Network Environment

Mary Anne Scott
scott@er.doe.gov
301-903-6368



Data Sources

Advanced Scientific Computing Research Program



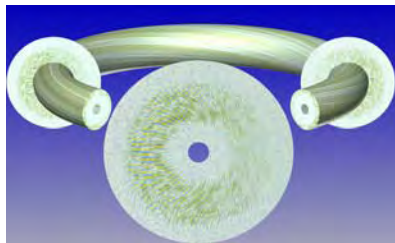
- Two different kinds of very large data sets:

- **Experimental data**

- High energy physics, environment and climate observation data, biological mass-spectrometry
- Data needs to be retained for long term

- **Simulation data**

- Astrophysics, climate, fusion, catalysis, QCD
- From computationally expensive large simulations
- Post processing of data using quantum Monte Carlo, analytics and graphical analysis, perturbation theory, and molecular dynamics





What is the Vision for the Future?

Advanced Scientific Computing Research Program

- Science, especially large-scale science, moves to a new regime that eliminates isolation, discourages redundancy, and promotes rapid scientific progress through the interplay of theory, simulation, and experiment.
- A seamless, high-performance network infrastructure in which science applications and advanced facilities are "n-way" interconnected to terascale computing, petascale storage, and high-end visualization capabilities.
- An advanced network facilitates collaborations among researchers and interactions between researchers and experimental and computational resources,



Program Objectives

Advanced Scientific Computing Research Program

- To develop a **scalable, secure, integrated, distributed** infrastructure to combine scientific resources and expertise across the Office of Science enterprise to address large-scale science projects that no single institution could undertake alone.
- To research, develop, test, and deploy advanced network technologies, scientific collaboration tools, and frameworks that enable scientists to integrate unique and expensive enterprise research facilities, computing resources, data archives, and expensive instruments into cost-effective virtual laboratories.



Planning Horizon for Network Environment Research

Advanced Scientific Computing Research Program

- Short term 1-3 yrs
 - Integration, prototyping, testing, and accelerated deployment of advanced computing, communications and middleware technologies
- Long term 1-5 yrs
 - Fundamental research issues for advanced collaborative and network capabilities addressing key issues for projected future applications requirements
- Keep the pipeline moving—research through development and prototyping



Strengths

Advanced Scientific Computing Research Program

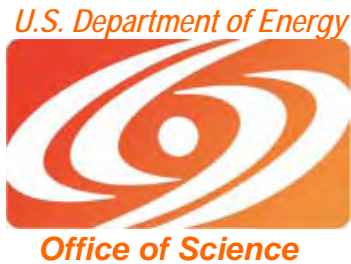
- **Fully integrated, real partnerships—application scientists, middleware developers, network researchers**
- **Lab leadership in distributed computing technology (grids)**
- **Effective pipeline from research through development**
- **Core groups developed over last ten years**
 - dedicated to program philosophy of collaboration/cooperation
- **Early adopters of technology in a production setting accelerates tool development and deployment**



Weaknesses

Advanced Scientific Computing Research Program

- **Cultural inertia**
 - Community building is hard
 - No mechanism for institutionalizing new services
 - Success metrics for production conflicts with success metrics for research and development
- **Technology barriers**
 - Perceived cost of adopting new technology vs. benefit
 - Lack of investment in fault-tolerance, error detection/recovery, etc
 - Limited ability to test at scale in a production network
 - Less than mature code with no support guarantee
- **Organizational barriers**
 - network Ineffective integration of program responsibilities and accountabilities moving from research to production
 - Minimal migration of new technologies into production



How is Knowledge Transferred

Advanced Scientific Computing Research Program

- **Partnerships between CS/network researchers and discipline scientists focused on specific problem areas drive the initial technology transfer**
- **Goal of supporting collaborative/cooperative work is incorporated into the program and leads to high level of interaction, integration, and cooperation**
- **Incorporate new services as they mature into ESnet**
- **Outreach**
 - Process involvement
 - Demonstrations
 - Workshops, conferences
 - Newsletters, reports
- **Support standards development**



Gap Analysis

Advanced Scientific Computing Research Program

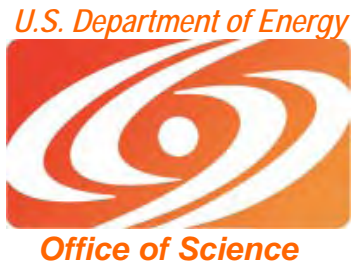
- **End-to-end performance**
 - Multi-domain
 - Ultra high-speed transport protocol
 - Network measurement and prediction
- **Cyber security**
 - Scalable distributed authentication and authorization systems
 - Ultra high-speed network components
- **High-Performance Middleware**
 - Network caching and computing
 - Real-time collaborative control and data streams
 - Fault-tolerance, error detection/correction
- **Integrated testbeds and networks**
 - Network research to accelerate advanced technologies
 - Experimental deployment of high-impact applications



Opportunities

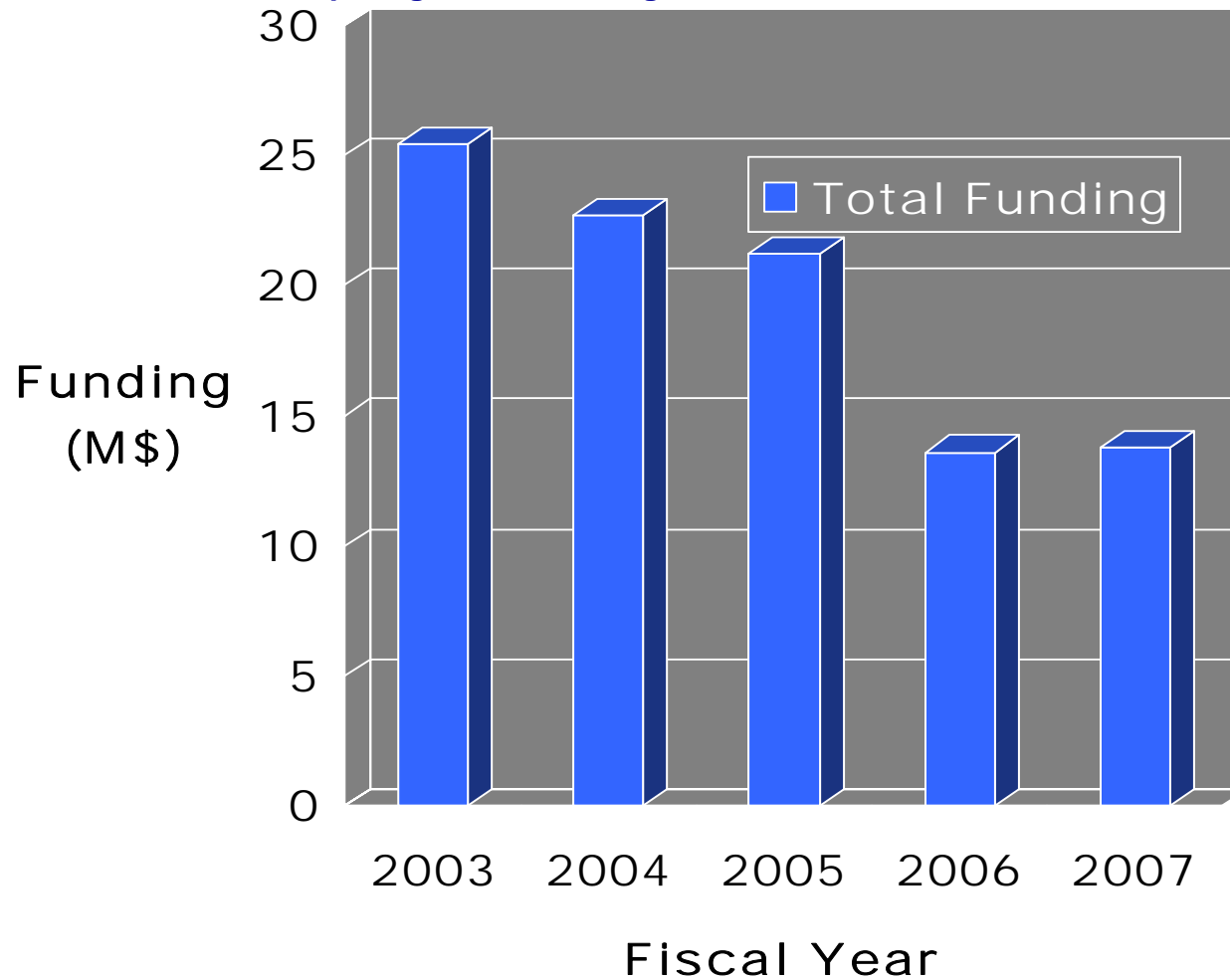
Advanced Scientific Computing Research Program

- Adopt abundant cost-effective capacity in the optical core networks to develop cost-effective agile network infrastructures to support high-impact science applications
- Develop gigabit networks and services for interconnecting data analysis and management centers associated with Petascale computers
- Partnerships between scientists and network/middleware researchers can be exploited to increase scale and productivity of science in areas like bioinformatics and nanotechnology
- Apply research network testbeds to accelerate the development of advanced networking technologies, especially approaches for operational cybersecurity
- Develop experimental network testbeds to test and validate new network technologies using real world applications



Distributed Network Environment Research Funding

Advanced Scientific Computing Research Program





Research Investments

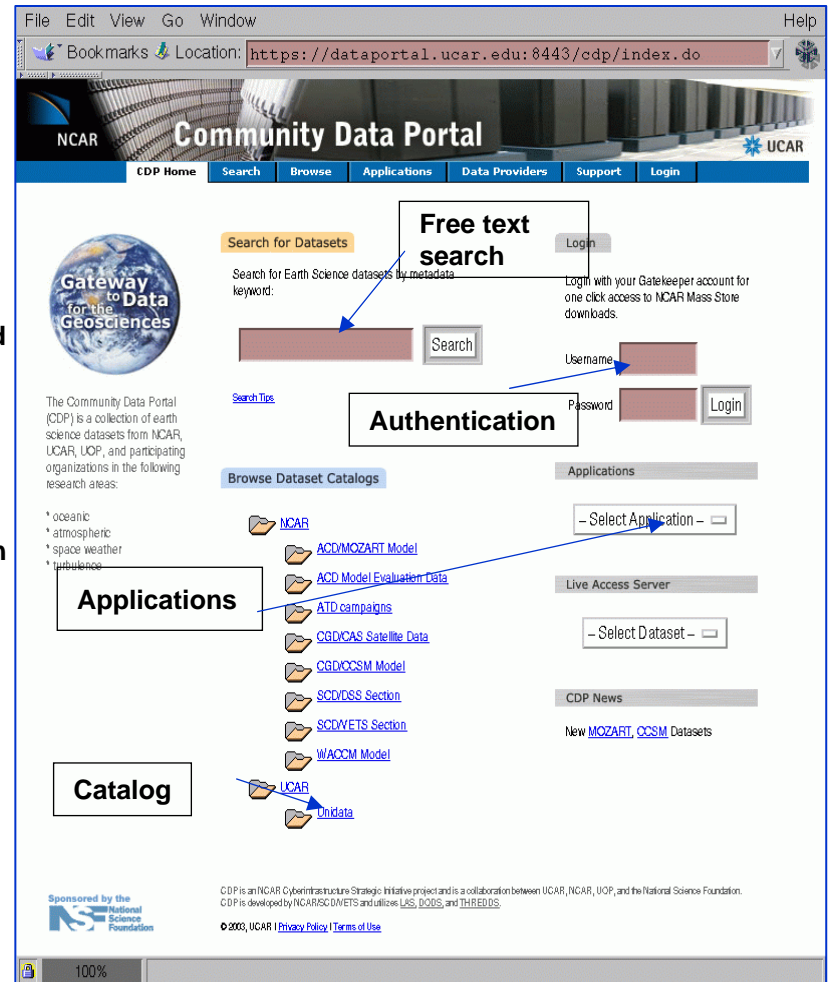
Advanced Scientific Computing Research Program

- **Middleware**
 - Facilitate the discovery and utilization of scientific data, computers, software, and instruments over the network in a controlled fashion
 - Integrate remote resources and collaboration capabilities into local experimental and computational environments
 - Services to support collaborative scientific work
- **Network Research**
 - Network measurement and analysis for understanding end-to-end performance
 - On-demand bandwidth
 - Security for open environments
- **Pilots and Research Testbeds**
 - Early implementations of virtual laboratories and experimental infrastructure to test and validate the enabling network and middleware technologies for large-scale science applications under realistic conditions
 - Unite distributed expertise, instruments, and computers for discipline-specific applications

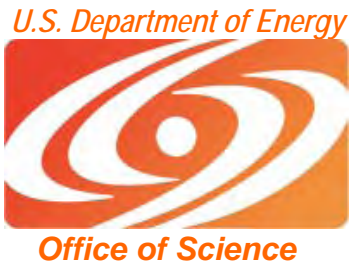
Enabling Large-scale Science Earth System Grid (ESG)

Advanced Scientific Computing Research Program

- **Payoffs of this distributed collaborative infrastructure include**
 - Distributed data-sharing
 - Simplified data discovery of climate data
 - Large-scale climate data processing and analysis
 - Increased collaboration among climate research scientists
 - Aid in climate assessments and estimates of future climate variability and trends
- **Two Portals are serving climate data**
 - Nearly 100 terabytes of simulations from the latest version of Community Climate System Model (CCSM) under various emission scenarios are available
 - Output from the 16 models (~30 TB) available to scientists participating in the Intergovernmental Panel for Climate Change (IPCC) Working Group I
 - data served to about 400 projects for over ten months
 - average data movement is about 300 GB per day
 - Over 200 research papers are in preparation
- **ESG integrates supercomputer with large-scale data and analysis services to create a powerful environment for next general climate research**



The screenshot shows the Community Data Portal (CDP) interface. At the top, there is a navigation bar with links for CDP Home, Search, Browse, Applications, Data Providers, Support, and Login. The main content area includes a search bar labeled 'Search for Datasets' and a 'Free text search' box. Below the search bar is a 'Search' button. To the right, there is a login section with fields for 'Username' and 'Password', and a 'Login' button. Below the login section is a 'Browse Dataset Catalogs' section with a list of datasets including 'ACD/MOZART Model', 'ACD Model Evaluation Data', 'ATD campaigns', 'CGD/CAS Satellite Data', 'CGD/CCSM Model', 'SCD/DSS Section', 'SCD/VETS Section', 'WACCM Model', and 'UCAR Unidata'. There is also an 'Applications' section with a dropdown menu for selecting an application. At the bottom, there is a 'Catalog' section. The page is sponsored by the National Science Foundation and includes a footer with copyright information and a privacy policy link.

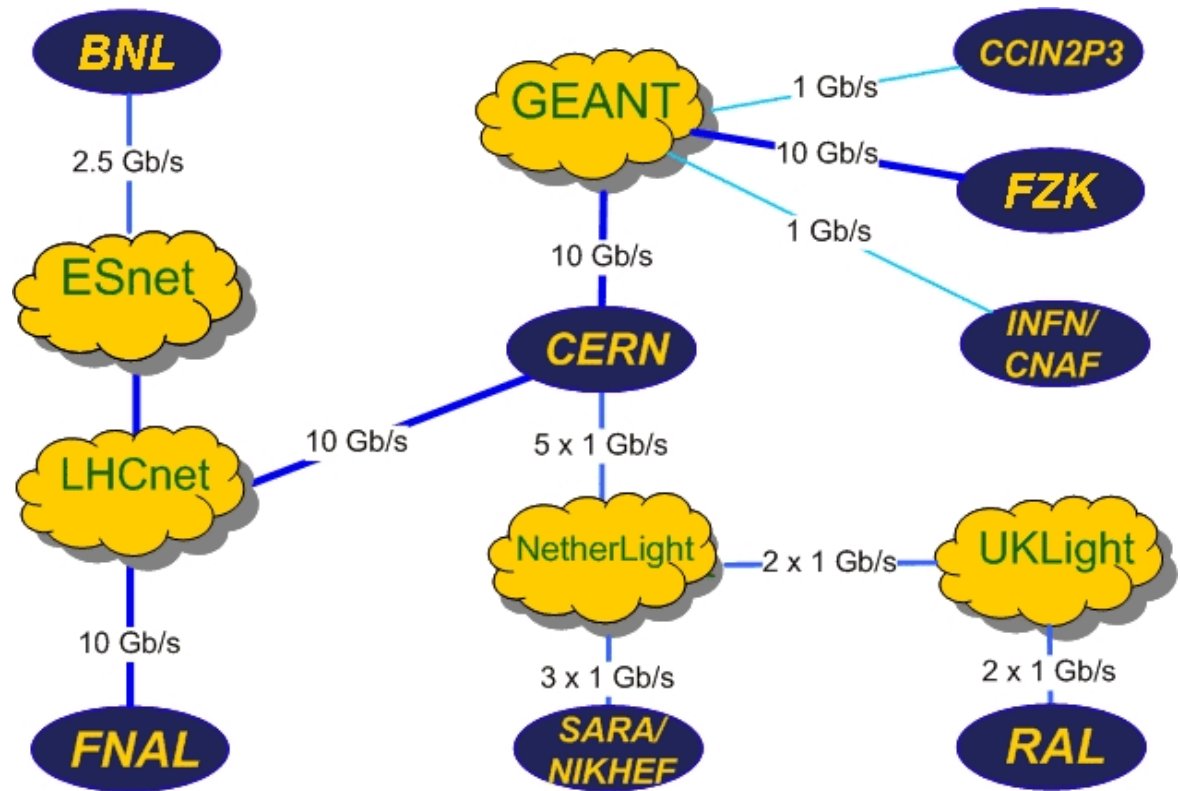


Middleware tool—a key enabler for a recent LHC “data challenge”

Advanced Scientific Computing Research Program

“Geneva, 25 April 2005 –

“Today, in a significant milestone for scientific grid computing, eight major computing centres successfully completed a challenge to sustain a continuous data flow of 600 megabytes per second (MB/s) on average for 10 days from CERN1 in Geneva, Switzerland to seven sites in Europe and the US. The total amount of data transmitted during this challenge—500 terabytes—would take about 250 years to download using a typical 512 kilobit per second household broadband connection.”



Every byte of that 500 terabytes was moved with GridFTP, developed as part of a SciDAC DataGrid Middleware project—enabling science on a scale impossible without such tools.