

# CEDPS

## Center for Enabling Distributed Petascale Science

---

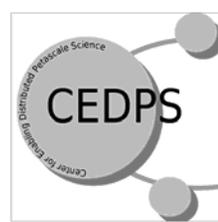
Jennifer M. Schopf

[jms@mcs.anl.gov](mailto:jms@mcs.anl.gov)

Argonne National Laboratory

- General Intro to Distributed Systems, Grids, and CEDPS
- Data Management Tools
  - Ann Chervenak, ISI
- Troubleshooting Tools
  - Brian Tierney, LBNL
- Question and Answer Session

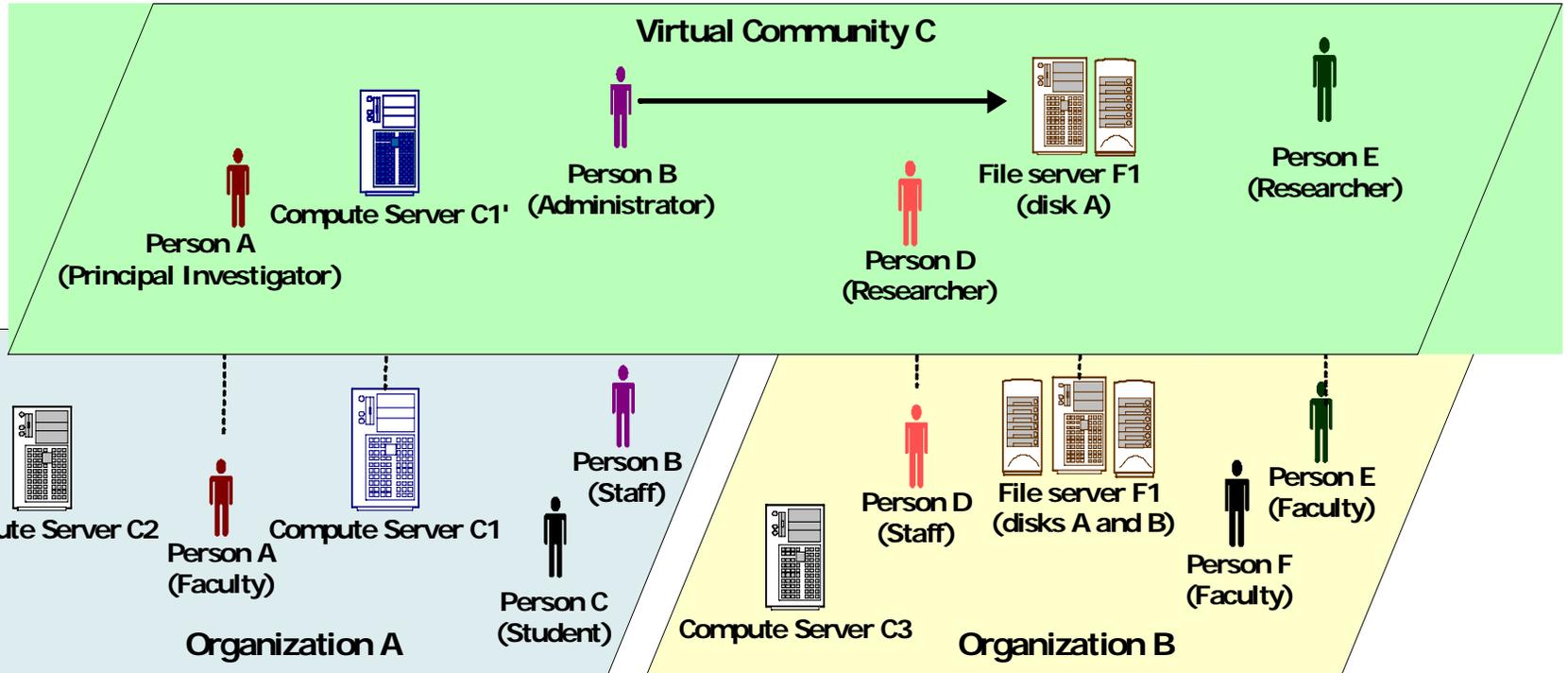
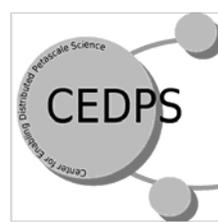
# What is a Grid?



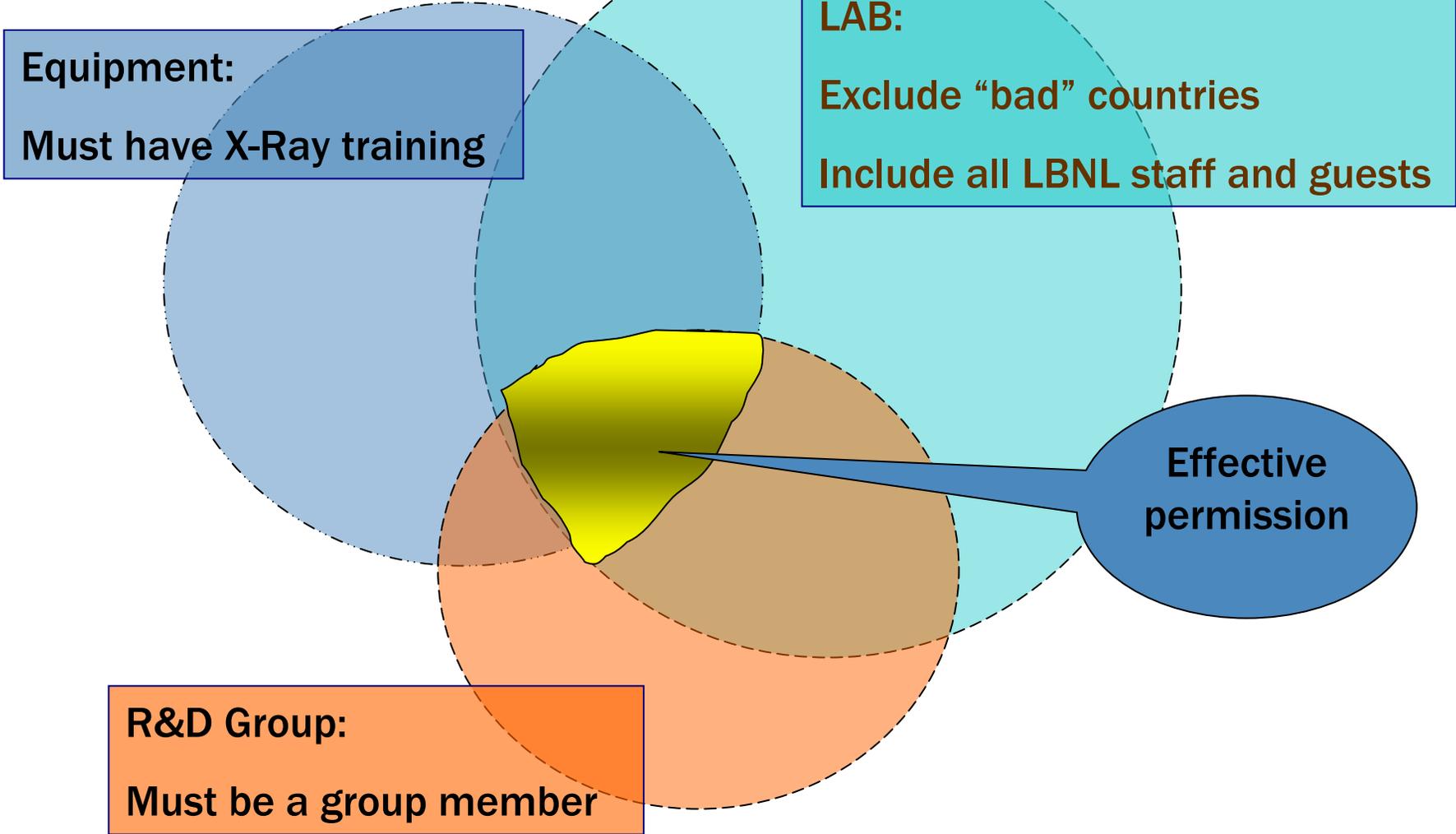
- Resource sharing
  - Computers, storage, sensors, networks, ...
  - Sharing always conditional: issues of trust, policy, negotiation, payment, ...
- Coordinated problem solving
  - Beyond client-server: distributed data analysis, computation, collaboration, ...
- Dynamic, multi-institutional virtual orgs
  - Community overlays on classic org structures
  - Large or small, static or dynamic



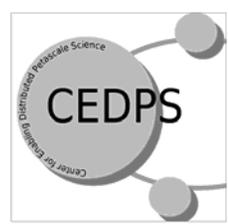
# Virtual Organization (VO) Concept



- VO for each application or workload
- Carve out and configure resources for a particular use and set of users

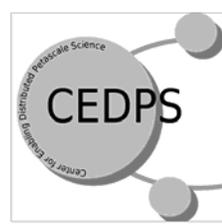


# Why Are Grids Hard?



- Lack of central control
  - Where things run
  - When they run
- Shared resources
  - Contention, variability
- Communication and coordination
  - Different sites implies different sys admins, users, institutional goals, and often socio-political constraints

# So Why Do It?



- Computations that need to be done with a time limit
- Data that can't fit on one site
- Data owned by multiple sites
  
- Applications that need to be run bigger, faster, more



If planes had  
sped up by the  
same factor as  
computers over  
the past 50 years,  
we would cross  
the country in a  
tenth of a second

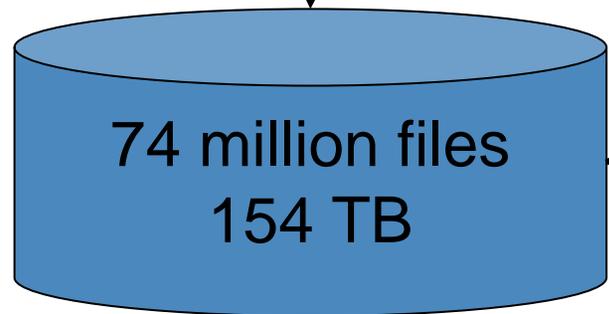
A photograph of a heavily congested city street, likely in New York City, showing a dense traffic jam. The street is filled with cars, taxis, and a yellow school bus. In the background, there are tall brick buildings and streetlights. A white speech bubble with a black border is overlaid on the right side of the image, containing the text: "Yes, but it would still take us two hours to get downtown!!!".

Yes, but it  
would still take  
us two hours to  
get downtown!!!

# FLASH Turbulence Simulation (Robert Fisher, Don Lamb, et al.)

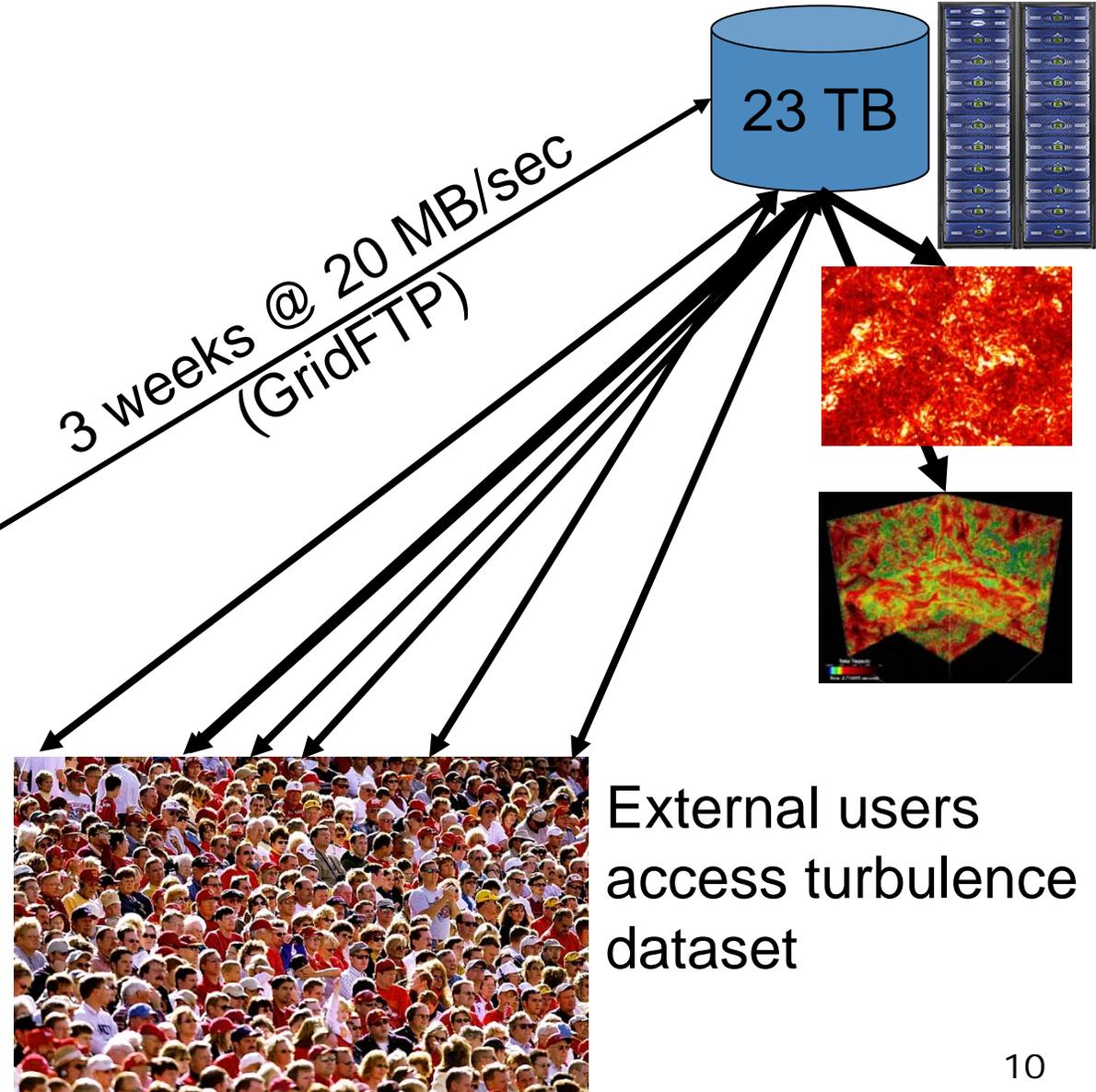


THE UNIVERSITY OF CHICAGO



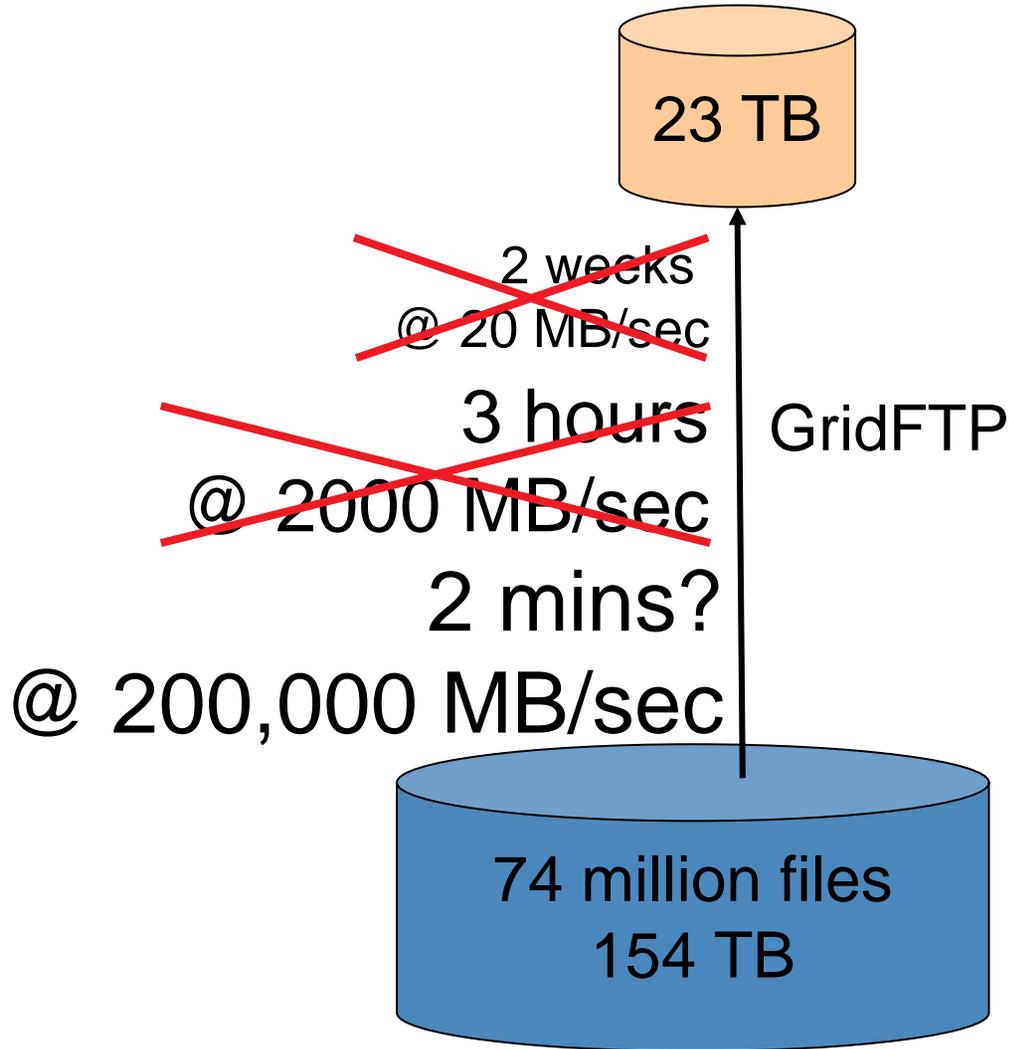
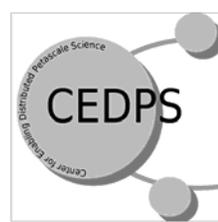
Largest compressible homogeneous isotropic turbulence simulation

Slide Courtesy of Ian Foster

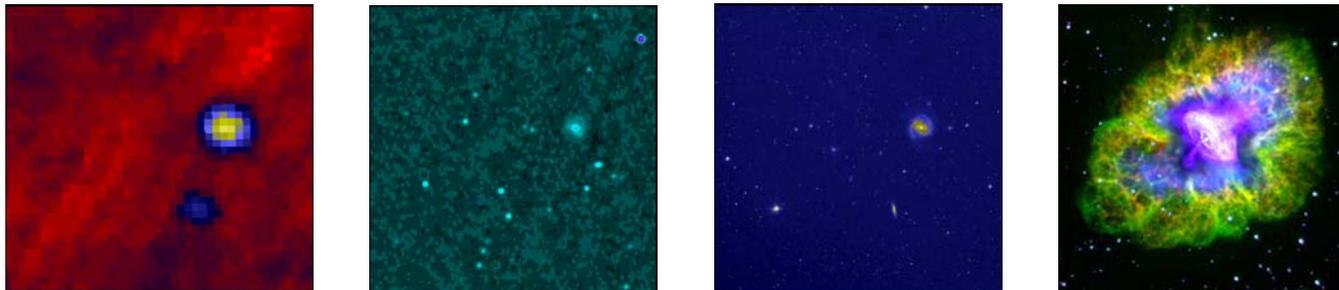




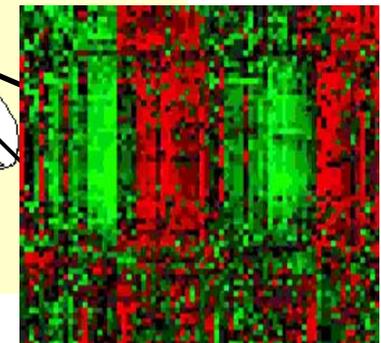
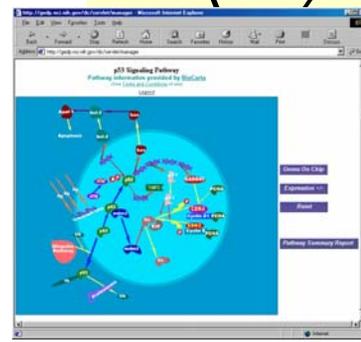
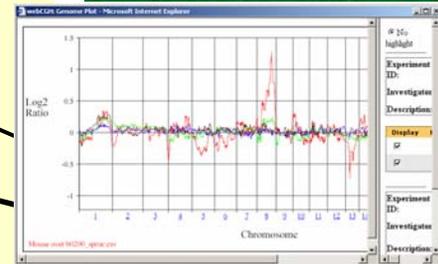
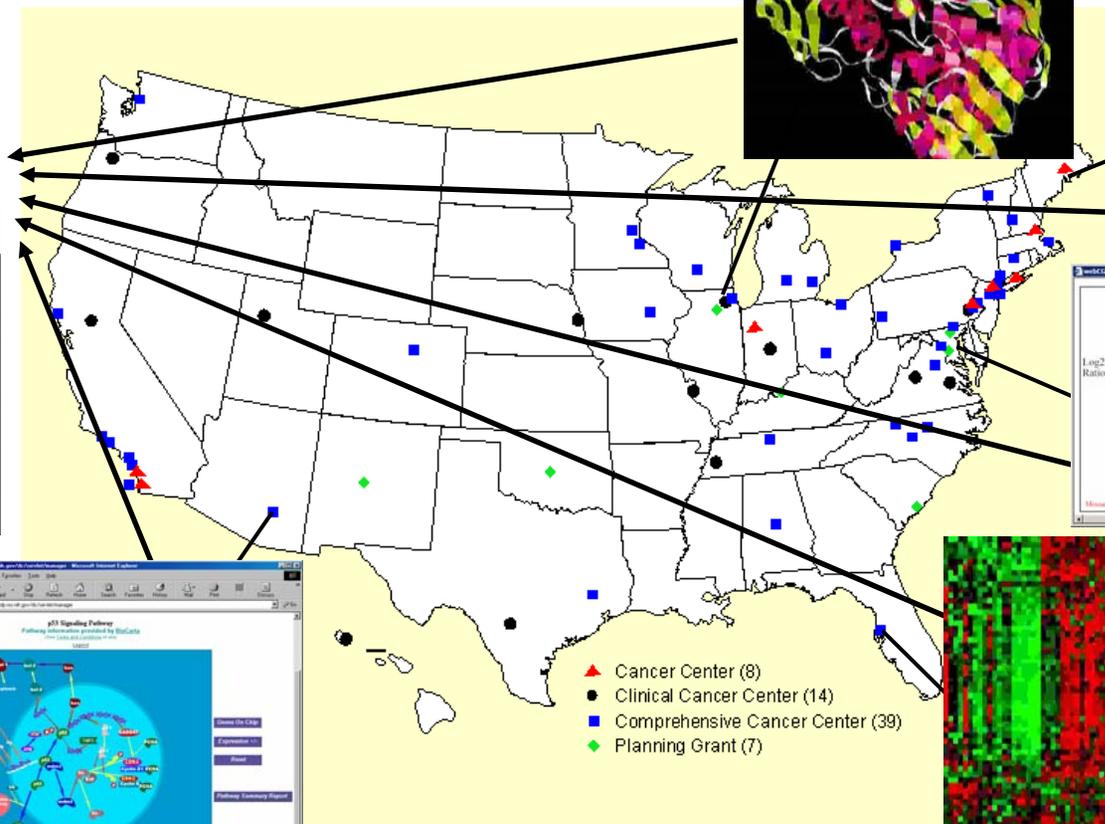
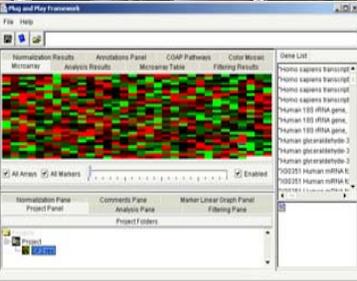
# Data Delivery Challenges



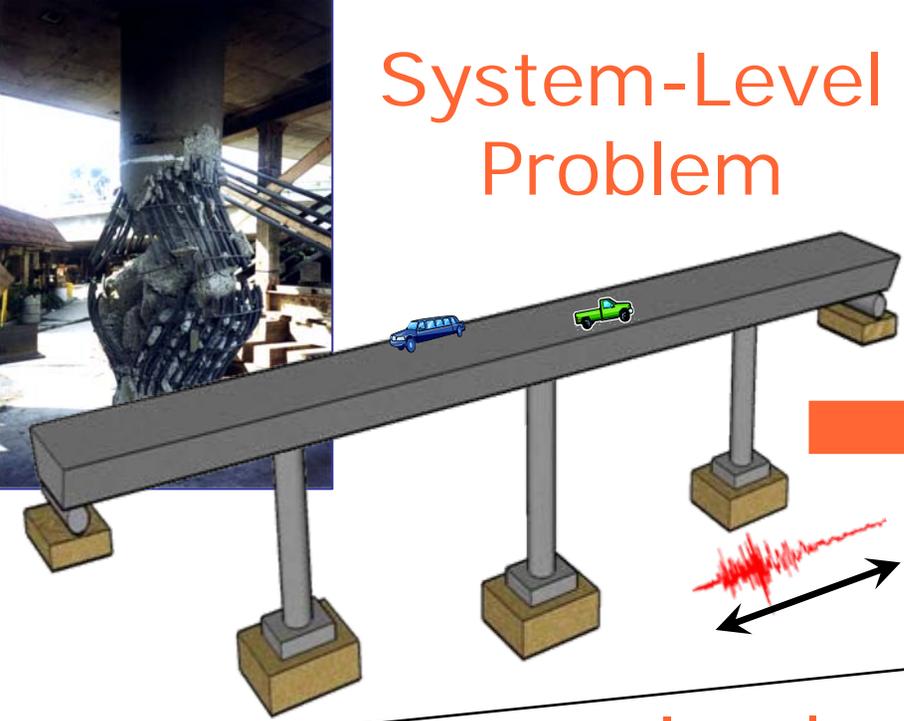
- Digital observatories provide online archives of data at different wavelengths



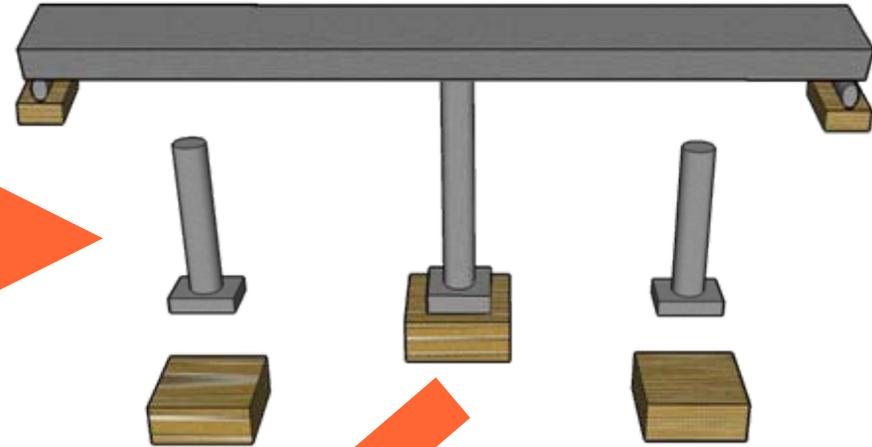
- Ask questions such as: what objects are visible in infrared but not visible spectrum?



# System-Level Problem

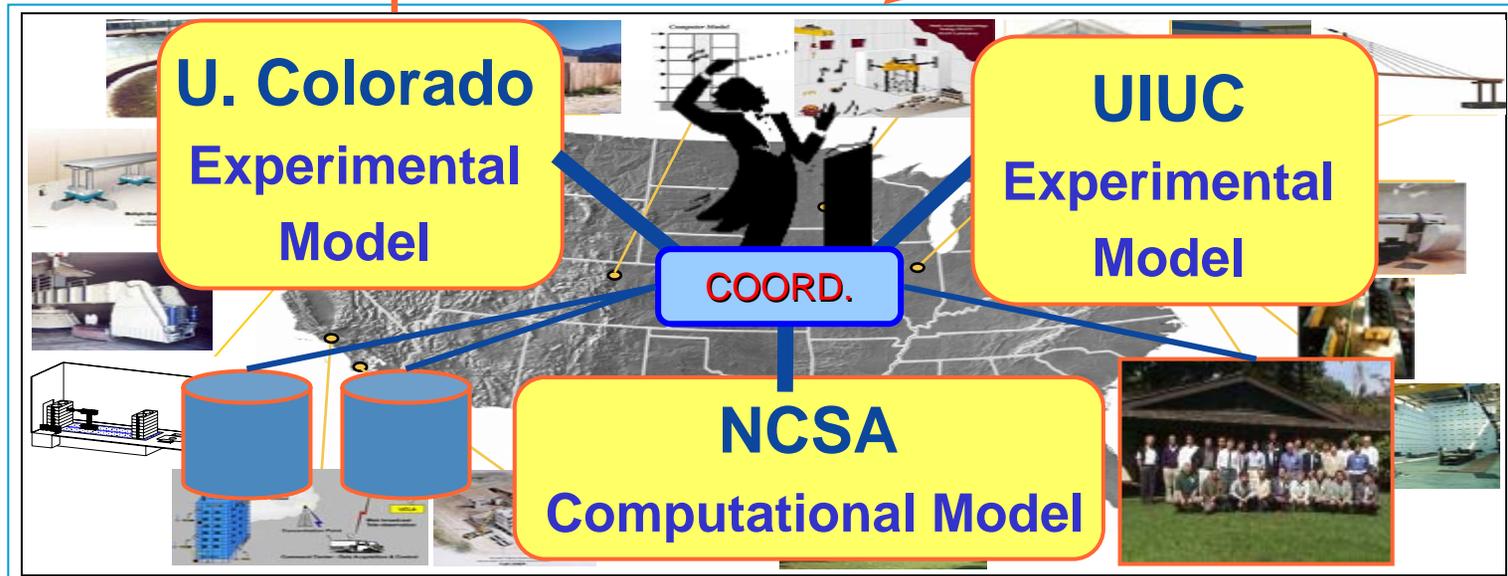


# Decomposition

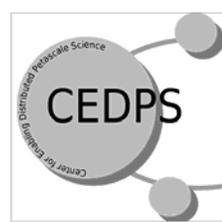


# Implementation

Facilities  
Computers  
Storage  
Networks  
Services  
Software  
People

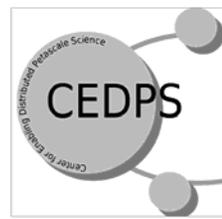


# What Kinds of Applications?



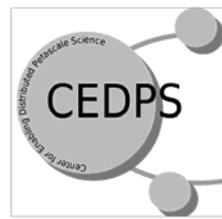
- Computation intensive
  - Interactive simulation (climate modeling)
  - Large-scale simulation and analysis (galaxy formation, gravity waves, event simulation)
  - Engineering (parameter studies, linked models)
- Data intensive
  - Experimental data analysis (e.g., physics)
  - Image & sensor analysis (astronomy, climate)
- Distributed collaboration
  - Online instrumentation (microscopes, x-ray)
  - Remote visualization (climate studies, biology)
  - Engineering (large-scale structural testing)

# Key Common Features



- The size and/or complexity of the problem
- Collaboration between people in several organizations
- Sharing computing resources, data, instruments

# A Grid is not...

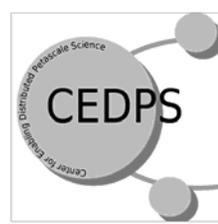


- SETI@home / BOINK
- FightAidsAtHome
- distributed.net
- File sharing systems like Kazaa
- Batch schedulers, cluster managers, and storage systems that happen to be connected to the Internet

# Elements Include ...

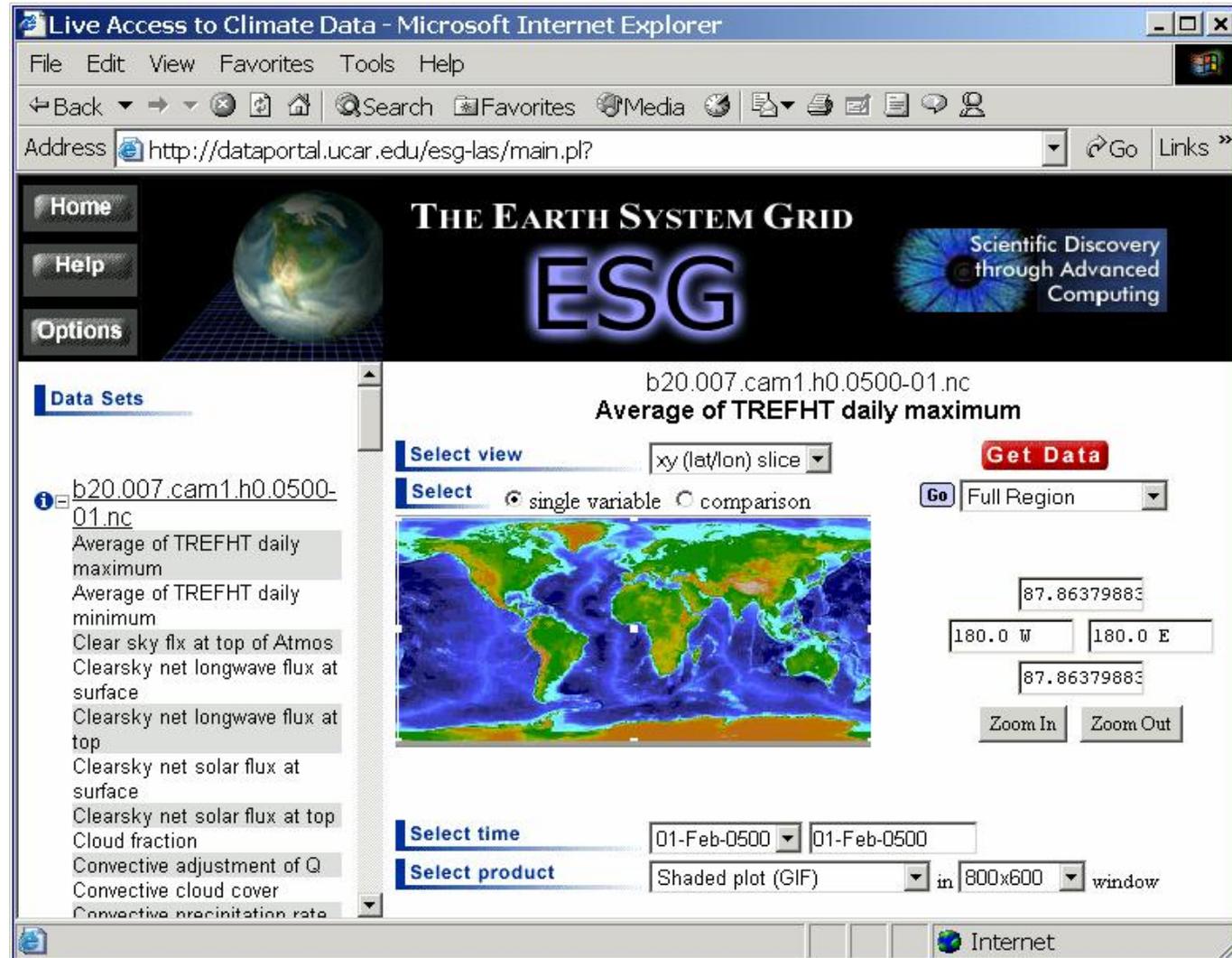
- Massively parallel petascale simulation
- High-performance parallel I/O
- Remote visualization
- High-speed reliable data movement
- Terascale local analysis
- Data access and analysis by external users
- Troubleshooting problems in end-to-end system
- Security
- Orchestration of these various activities

# Grid Infrastructure Offers:



- Distributed management
  - Of physical resources
  - Of software services
  - Of communities and their policies
- Unified treatment
  - Build on Web services framework
  - Use WS-RF, WS-Notification (or WS-Transfer/Man) to represent/access state
  - Common management abstractions & interfaces

Goal: Enable sharing & analysis of high-volume data from advanced earth system models



Live Access to Climate Data - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail News Chat

Address <http://dataportal.ucar.edu/esg-las/main.pl?> Go Links

Home Help Options

THE EARTH SYSTEM GRID  
**ESG**  
Scientific Discovery through Advanced Computing

b20.007.cam1.h0.0500-01.nc  
Average of TREFHT daily maximum

Select view xy (lat/lon) slice **Get Data**

Select  single variable  comparison Go Full Region

87.86379883  
180.0 W 180.0 E  
87.86379883  
Zoom In Zoom Out

Select time 01-Feb-0500 01-Feb-0500

Select product Shaded plot (GIF) in 800x600 window

Internet

# ESG

## Facts and Figures

### ESG Portal at NCAR

130 TB of data at four locations

- 840,331 files
- Includes the past 6 years of joint DOE/NSF climate modeling experiments

3,200 registered users

Downloads to date

- 25 TB
- 91,000 files



Worldwide ESG user base

### IPCC AR4 ESG Portal

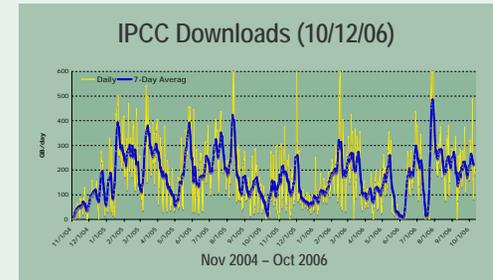
28 TB of data at one location

- 68,400 files
- Generated by a modeling campaign coordinated by the Intergovernmental Panel on Climate Change
- Model data from 11 countries

818 registered analysis projects

Downloads to date

- 123 TB
- 543,500 files
- 300 GB/day (average)



300 scientific papers published to date based on analysis of IPCC AR4 data

# Underlying Technologies

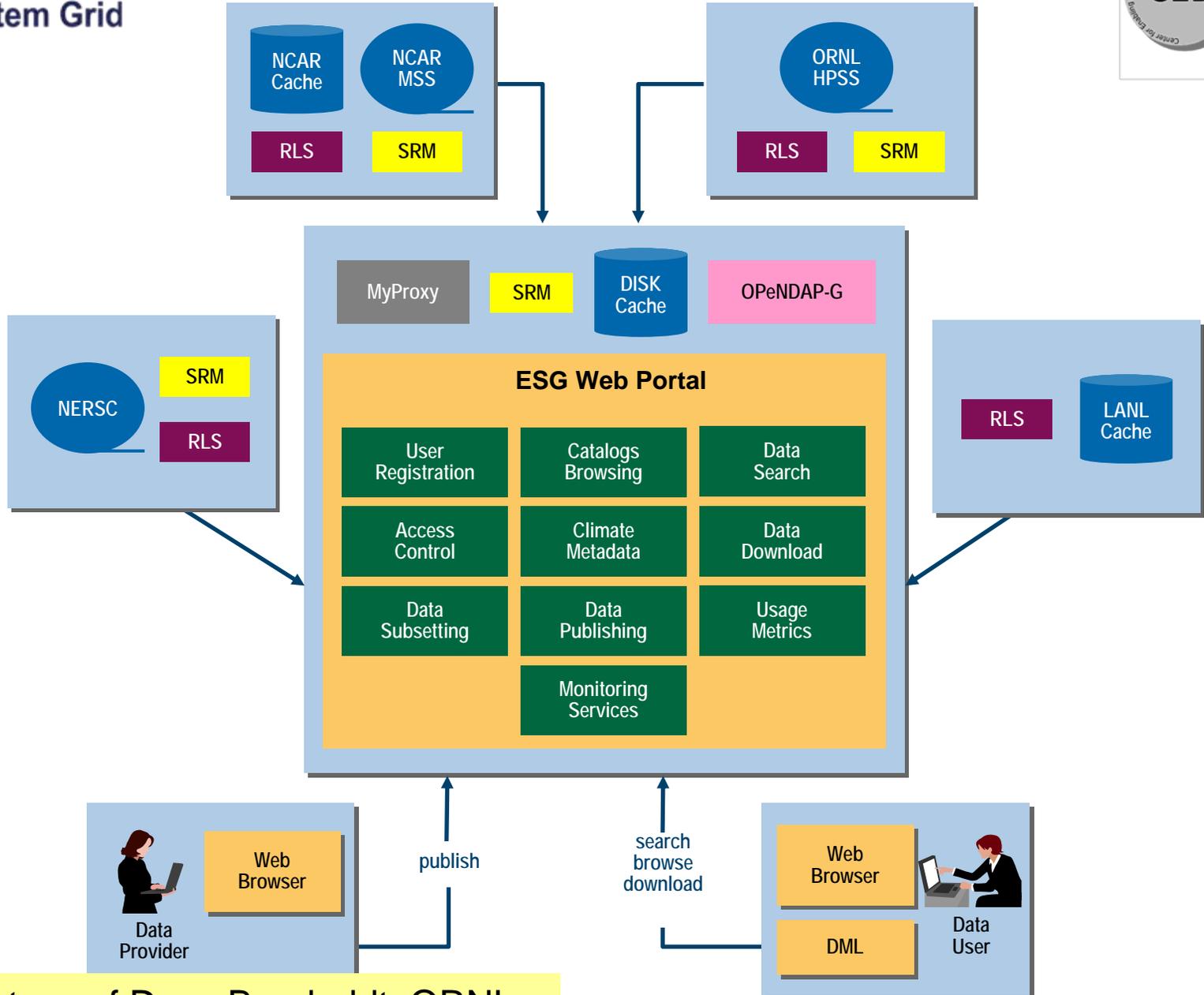
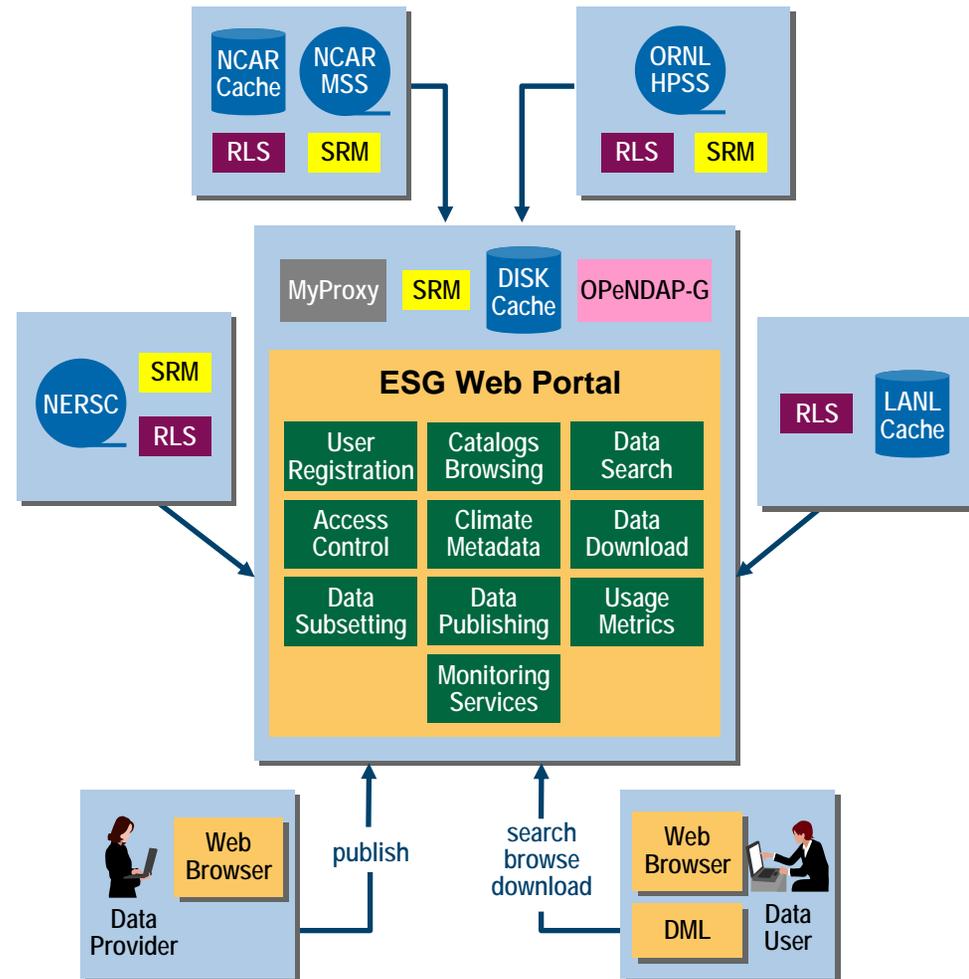
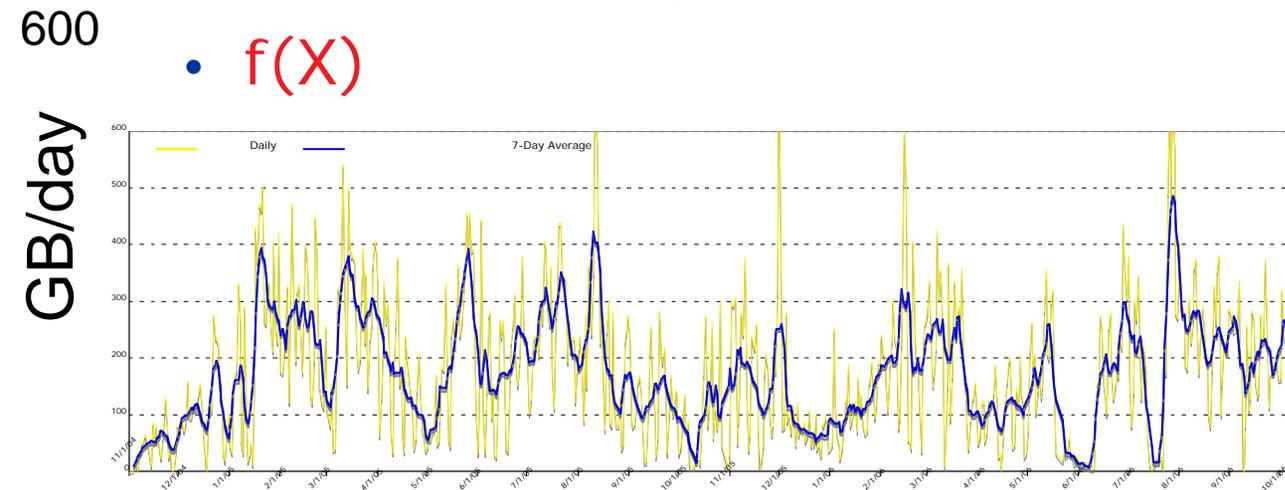
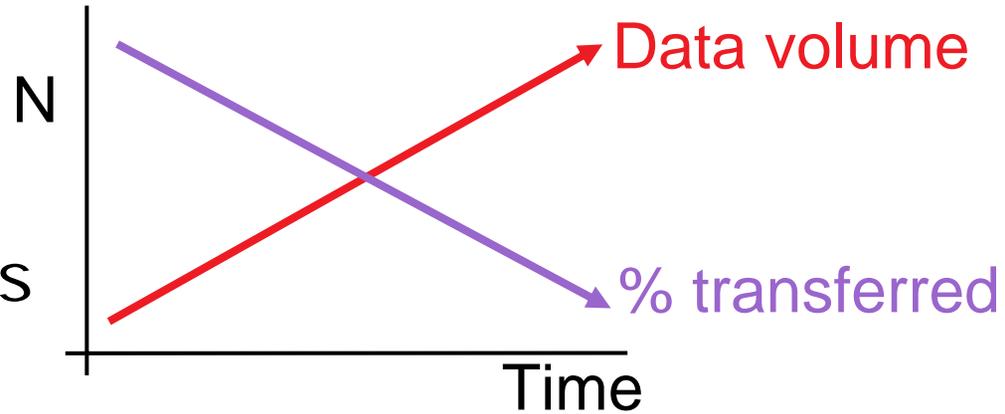


Figure Courtesy of Dave Bernholdt, ORNL

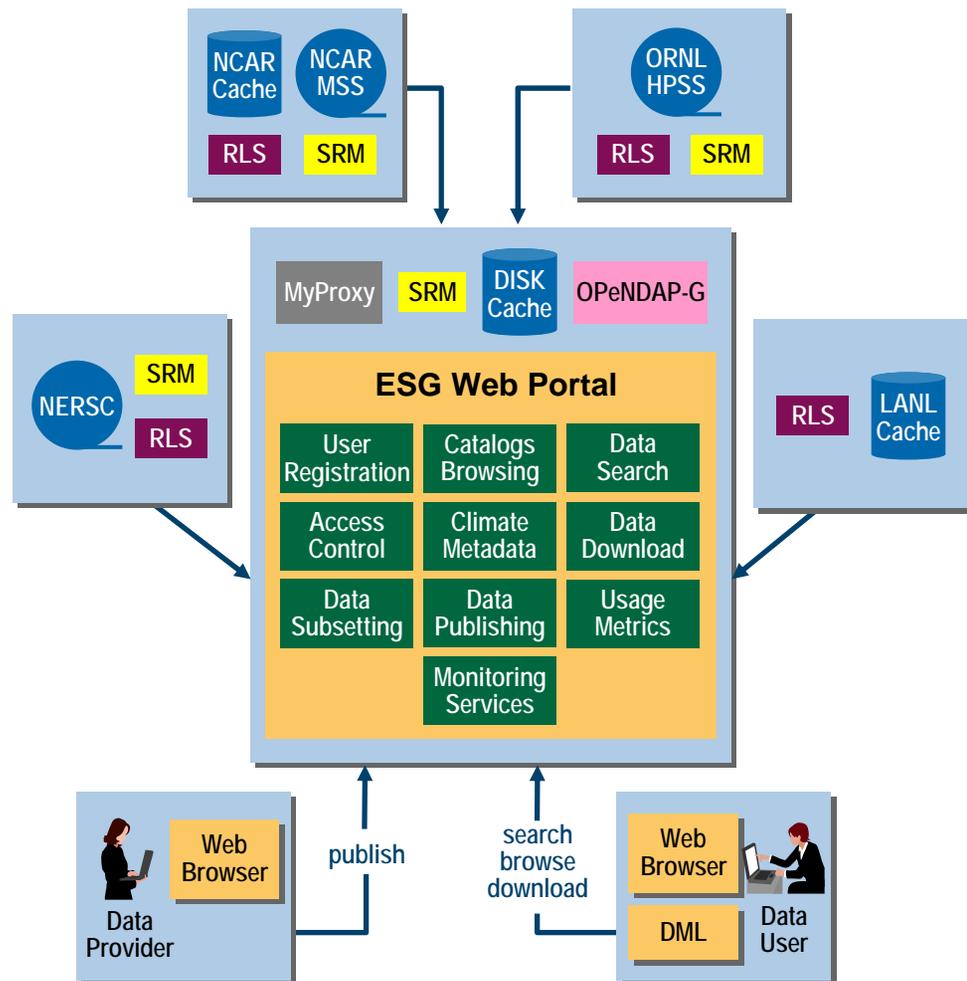
- Storage Resource Manager (SRM)
- Data Mover Lite (DML)
- GridFTP
- Globus Replica Location Service (RLS)



- Entire datasets
  - $X$
- Data subsets
  - $X[1:10, 1:50:2, 6]$
- Predefined operations
  - $ZonalMean(X)$
- User-defined operations
  - $f(X)$

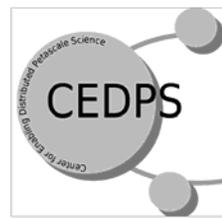


- Grid Security Infrastructure (GSI)
- MyProxy
- PURSE User registration





# Grids Do It Better: Monitoring ESG



- The climate community has come to depend on the ESG infrastructure as a critical resource
  - Failures of ESG components or services can disrupt the work of many scientists
  - Need to minimize infrastructure downtime
- Grid-wide monitoring needed in order to detect failures
  - Collect, aggregate, and sometimes act upon data describing system state

# Monitoring Overall System Status

- Monitored data are collected in Globus MDS4 Index service
- Information providers check resource status at a configured frequency
  - Currently, every 10 minutes
  - Report status to Index Service
- Index Service is queried by the ESG Web portal
  - Used to generate overall picture of state of ESG resources
  - Displayed on ESG Web portal page

**ESG Current Status**  
Updated: Tue Jun 27 16:52:32 MDT  
2006 MDT

	LANL	LBNL	NCAR	ORNL
MSS/MPSS		☹	☹	☹
SRM	☹	☹	☹	☹
FLS		☹	☹	☹
OpenDAPg			☹	
GridFTP server			☹	
HTTP server	☹		☹	

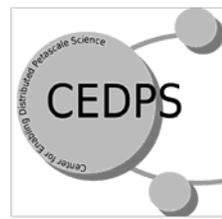
*(Explanation of current status)*

## Trigger Actions Based on Monitoring Information

- Globus MDS4 Trigger service periodically polls for data
- Based on the current resource status, Trigger service determines whether specified trigger rules and conditions are satisfied
  - Performs specified action for each Trigger, eg. send email to system administrators when services fail
  - System failures can be detected and corrected before they affect larger ESG community
- Future plans: include richer recovery operations as trigger actions, e.g., automatic restart of failed services



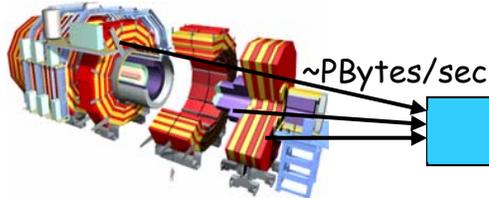
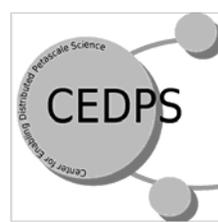
# Lessons Learned From ESG Monitoring



- Monitoring has significantly reduced downtime of ESG resources
  - Notifications allow failures to be addressed quickly
- Provides overview of current system state for users and system administrators
- Supports validation of new software deployments
- The monitoring system can be used to deduce reason for complex failures
  - System-wide monitoring can be used to detect a pattern of failures that occur close together in time
  - Deduce a problem at a different level of the system



# Data Grids for High Energy Physics



~PBytes/sec

Online System

~100 MBytes/sec

1 TIPS is approximately 25,000  
SpecInt95 equivalents

Offline Processor Farm  
~20 TIPS

~100 MBytes/sec

There is a "bunch crossing" every 25 nsecs.  
There are 100 "triggers" per second  
Each triggered event is ~1 MByte in size

**Tier 1**

**Tier 0**

CERN Computer Centre



~622 Mbits/sec  
or Air Freight (deprecated)

France Regional Centre

Germany Regional Centre

Italy Regional Centre

FermiLab ~4 TIPS

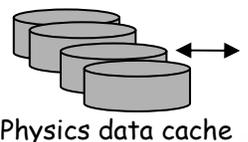


~622 Mbits/sec

**Tier 2**

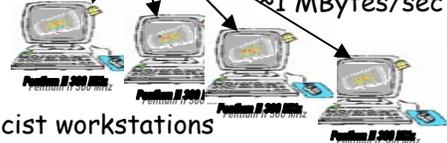
Caltech ~1 TIPS  
Tier2 Centre ~1 TIPS  
Centre  
Centre  
Centre

~622 Mbits/sec



Institute ~0.25TIPS  
Institute  
Institute  
Institute

~1 MBytes/sec

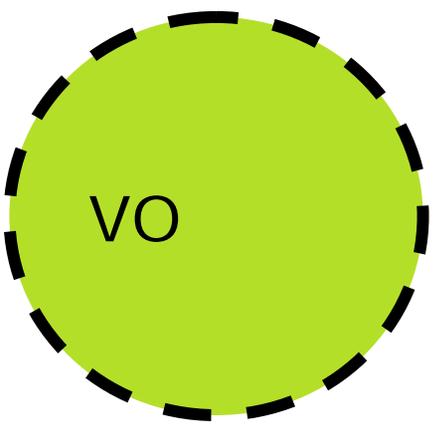
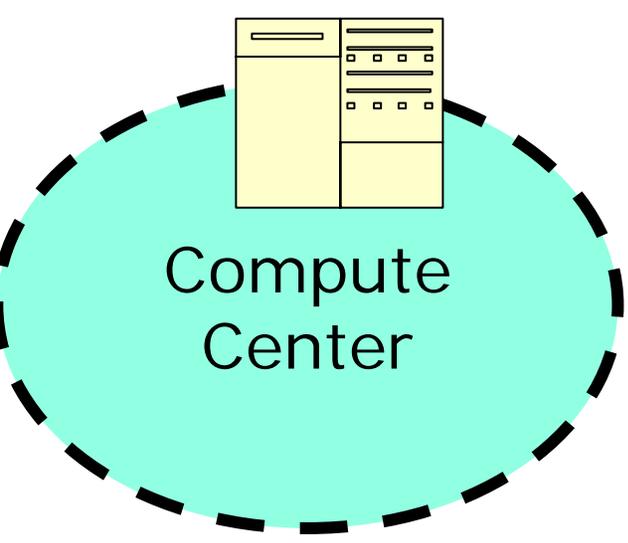


**Tier 4**

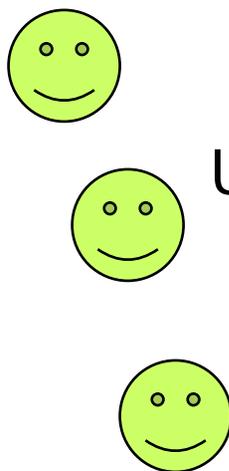
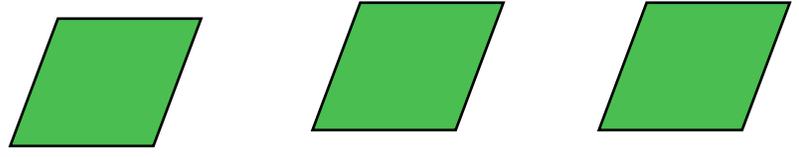
Physicists work on analysis "channels".  
Each institute will have ~10 physicists working on one or more channels; data for these channels should be cached by the institute server

Image courtesy Harvey Newman, Caltech

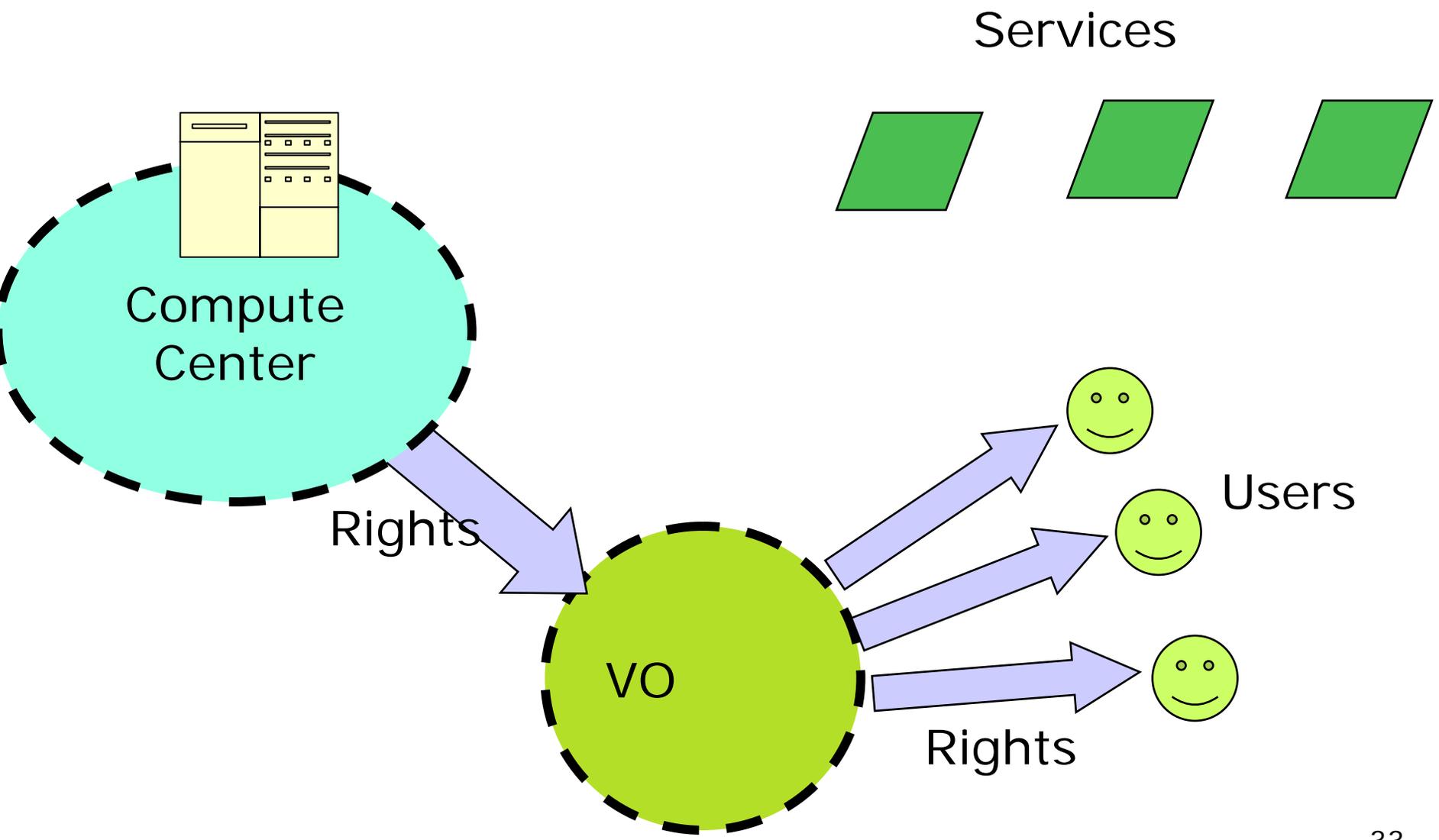
- Sites decide what data they want
  - Cross-site access policies must not over-ride local systems
- Users access data and compute resources
  - Use a metascheduler, which contacts lower level services on behalf of the user
  - Generic interfaces hide site heterogeneity from the end user

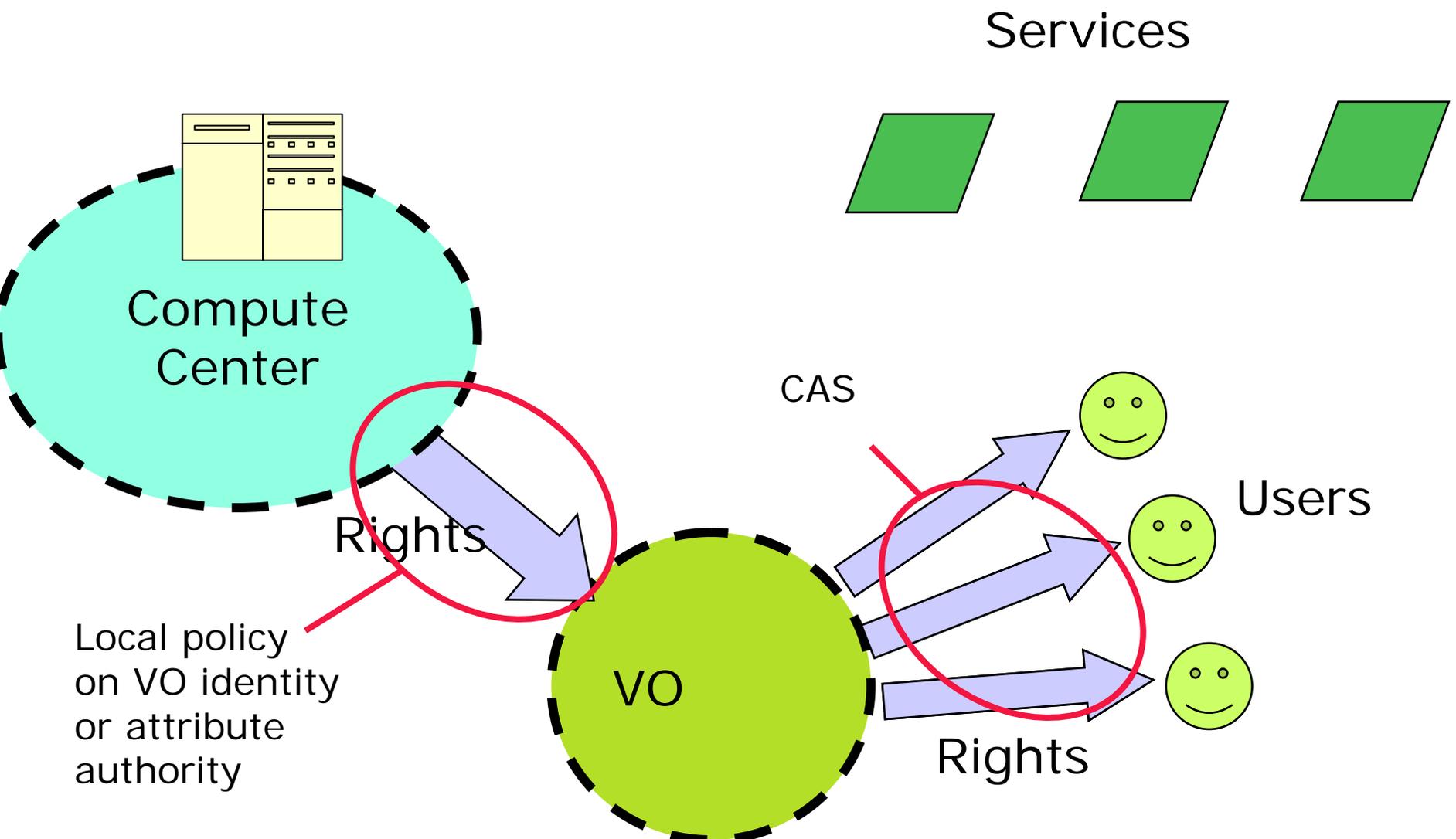


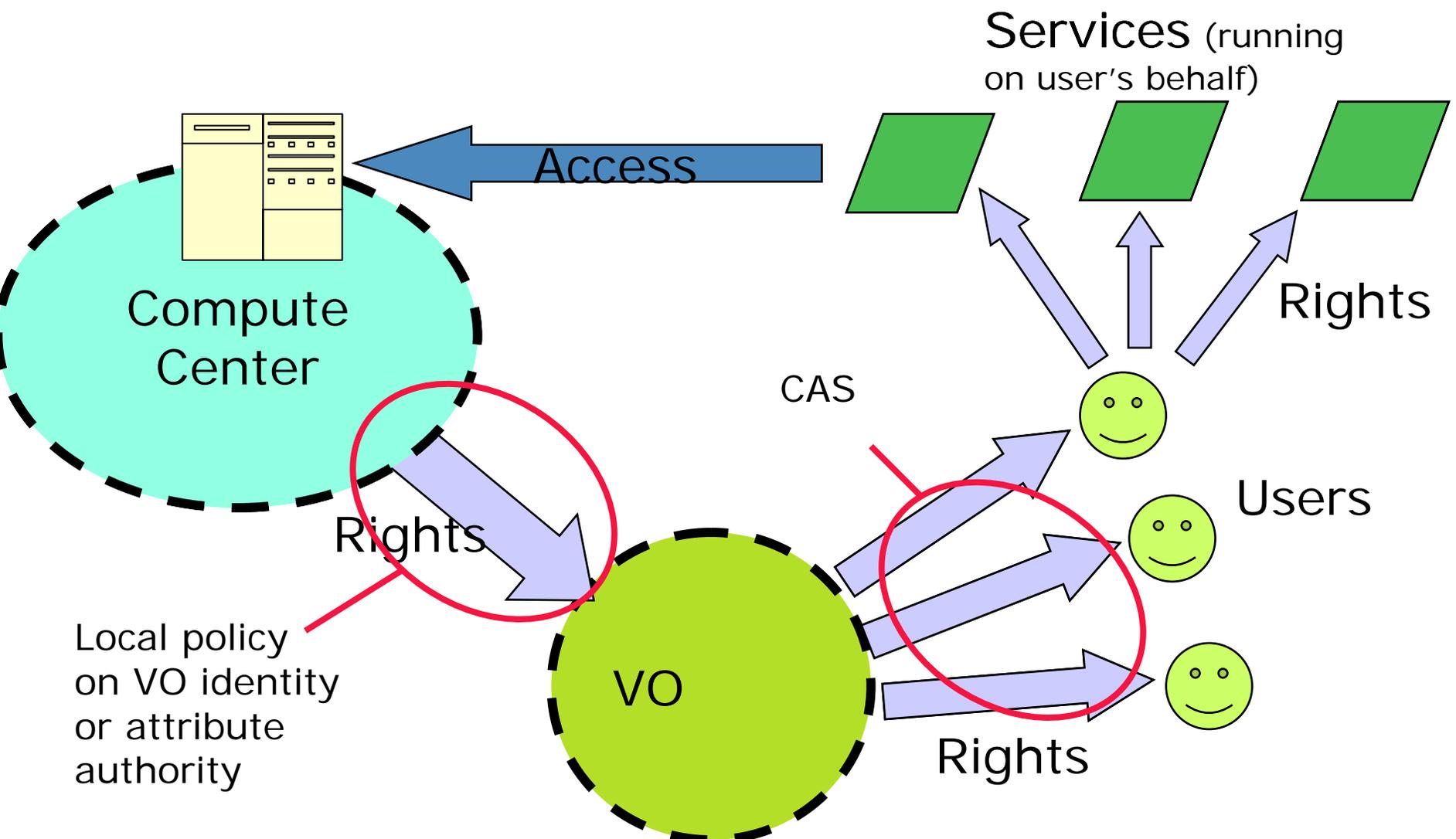
Services

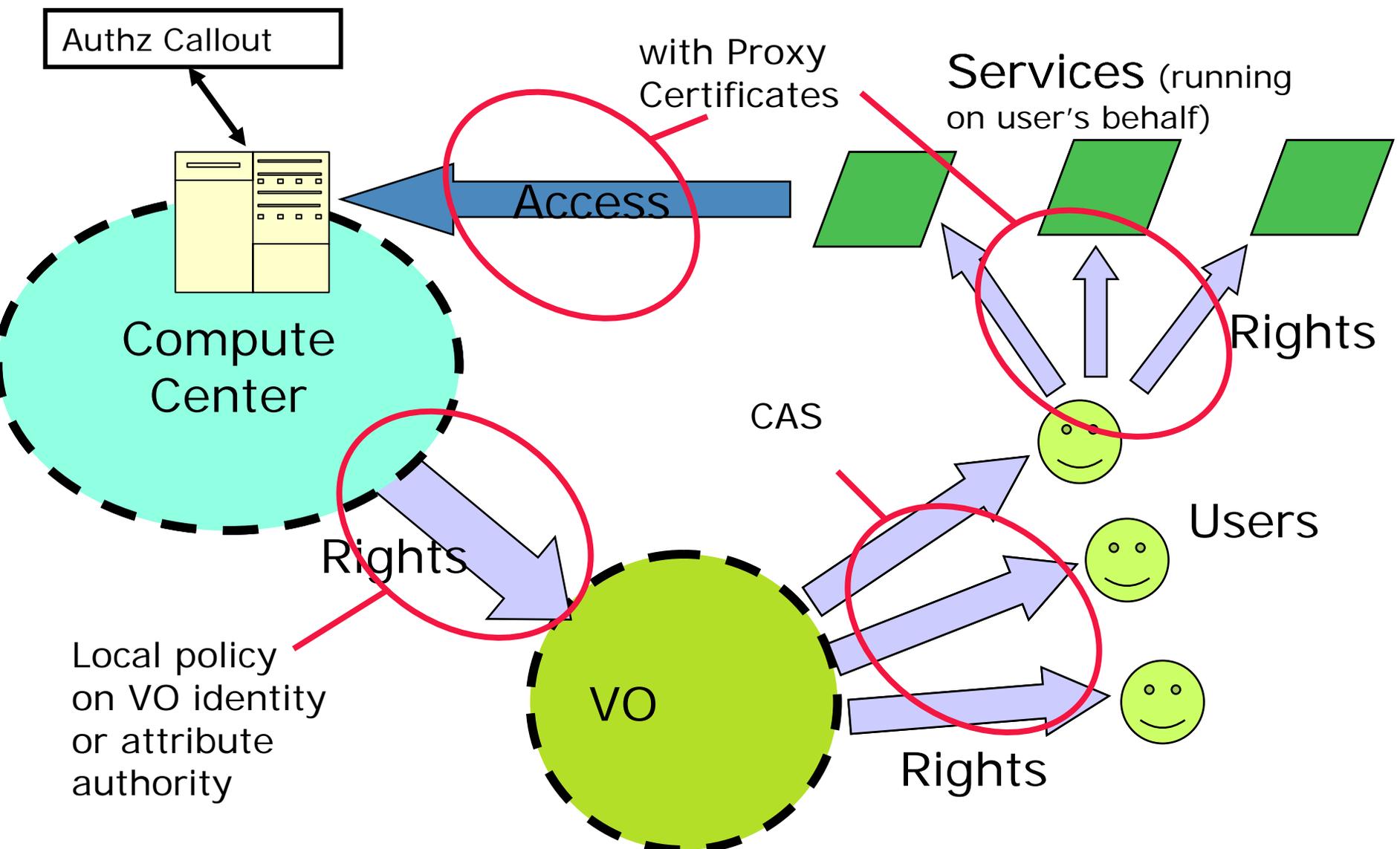


Users

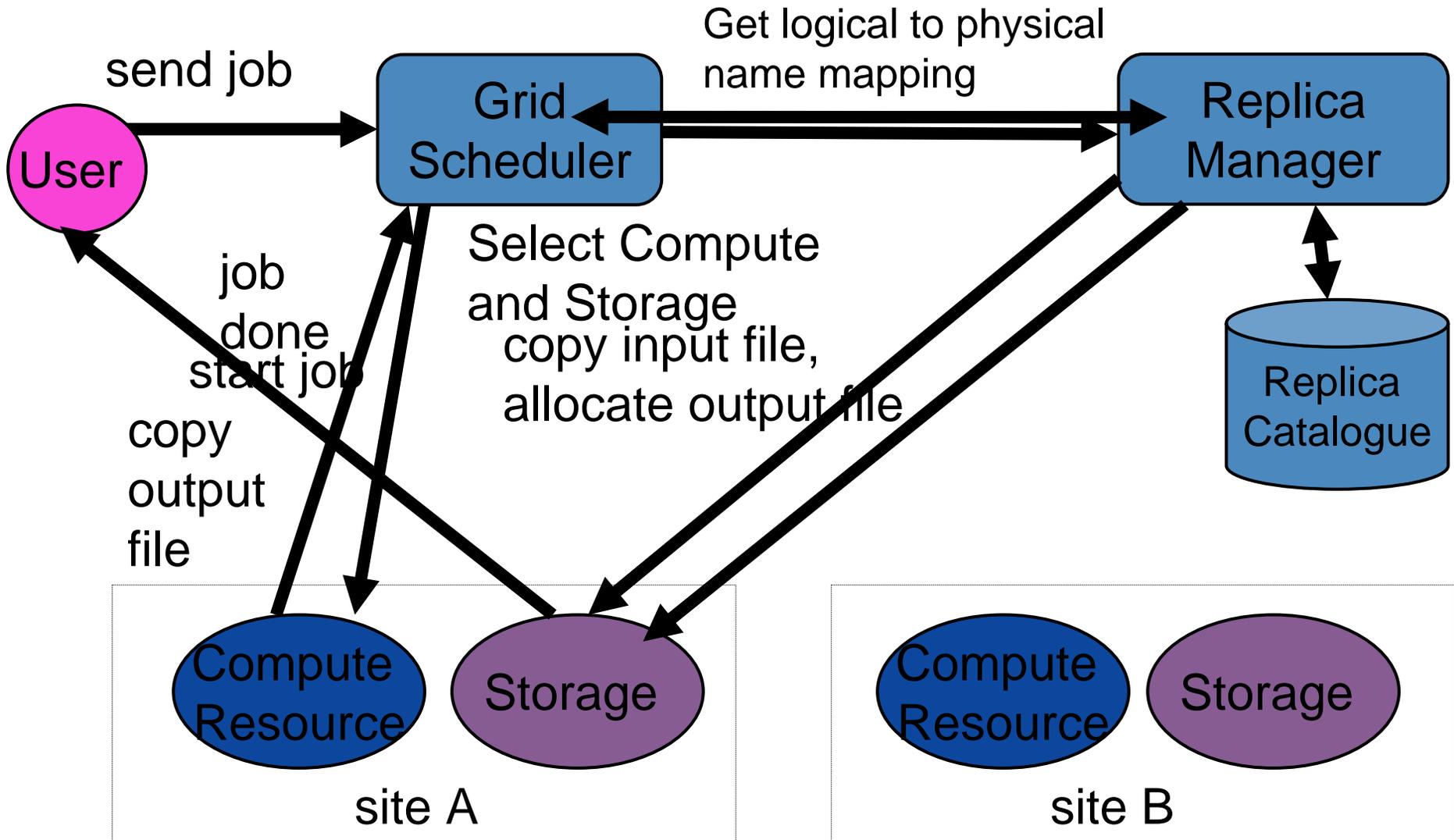








# Simple Grid Job Execution



- Any questions at this point?

- Center for Enabling Distributed Petascale Science
  - CEDPS – “seds” (silent P)
  - DOE SciDAC Center for Enabling Technology
  - July 1, 2006 – June 30, 2011, \$2.4M/yr
- Collaboration between 5 sites
  - Argonne National Laboratory
  - Fermi National Laboratory
  - Lawrence Berkeley National Laboratory
  - USC Information Sciences Institute
  - University of Wisconsin Madison
- Three focus areas
  - Moving data to compute resources
  - Moving compute services to data sites
  - Troubleshooting and diagnosis tools

- Massively parallel petascale simulation
- High-performance parallel I/O
- Remote visualization
- High-speed reliable data movement
- Terascale local analysis
- Data access and analysis by external users
- Troubleshooting problems in end-to-end system
- Security
- Orchestration of these various activities

- Massively parallel petascale simulation
- High-performance parallel I/O
- Remote visualization
- **High-speed reliable data mover**
- Terascale local analysis
- **Data access and analysis by external users**
- **Troubleshooting problems in end-to-end system**
- Security
- Orchestration of these various activities



Open Science Grid



Earth System Grid

# The Petascale Data Challenge

- DOE facilities generate **many petabytes** of data (2 petabytes = **all** U. S. academic research libraries!)

Remote distributed users

- **Remote users** (at labs universities, industry) need data!

- **Rapid, reliable access** key to maximizing value of \$B facilities

Massive data



DOE facilities

# Bridging the Divide (1): Move Data to Users When & Where Needed

"Deliver this 100 Terabytes to locations A, B, C by 9am tomorrow"

- **Fast:** >10,000x faster than usual Internet

- **Reliable:** recover from many failures
- **Predictable:** data arrives when scheduled
- **Secure:** protect expensive resources & data
- **Scalable:** deal with many users & much data

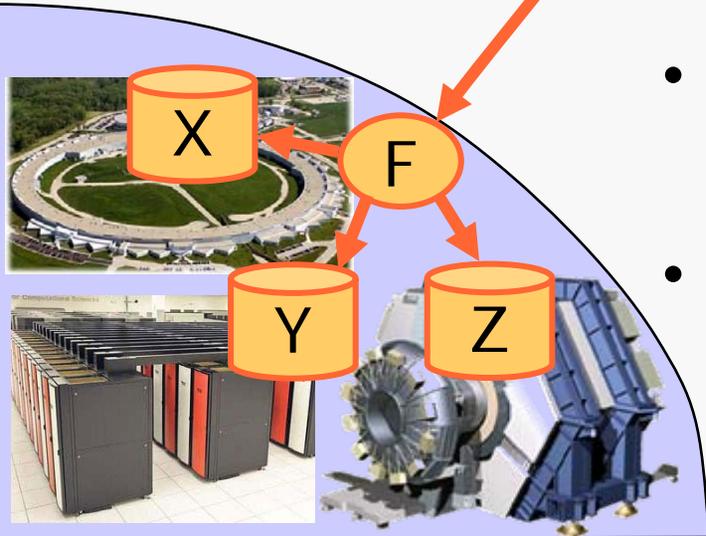


## *Bridging the Divide (2):* Allow Users to Move Computation Near Data

"Perform my  
computation F on  
datasets X, Y, Z"

- **Science services:** provide analysis functions near data source

- **Flexible:** easy integration of functions
- **Secure:** protect expensive resources & data
- **Scalable:** deal with many users & much data

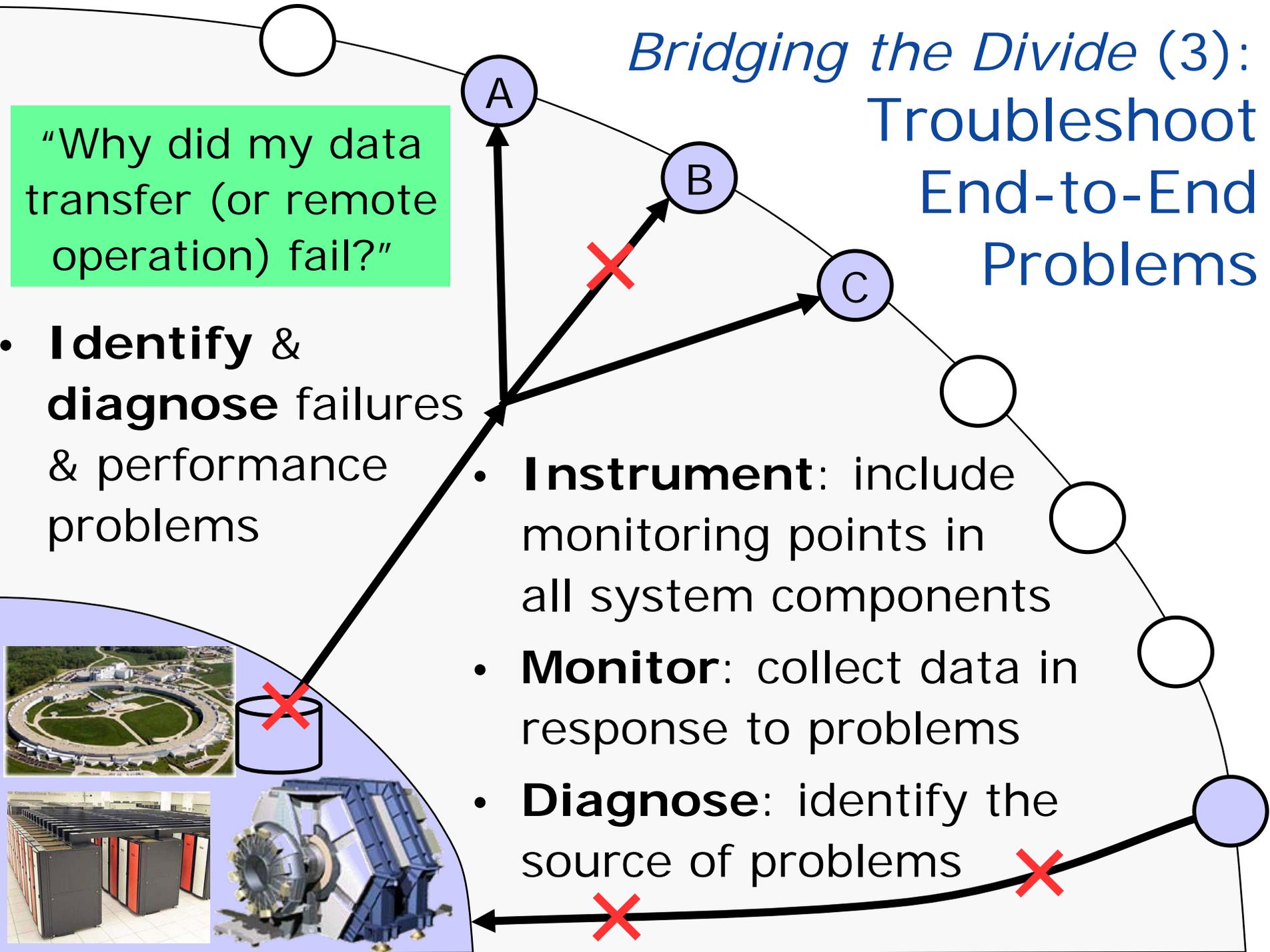


# Bridging the Divide (3): Troubleshoot End-to-End Problems

“Why did my data transfer (or remote operation) fail?”

- **Identify & diagnose** failures & performance problems

- **Instrument:** include monitoring points in all system components
- **Monitor:** collect data in response to problems
- **Diagnose:** identify the source of problems



- Jennifer Schopf
  - [jms@mcs.anl.gov](mailto:jms@mcs.anl.gov)
  - <http://www.mcs.anl.gov/~jms>
- CEDPS
  - <http://www.cedps.net>