GlobusWorld 2011:
When Grid meets cloud …

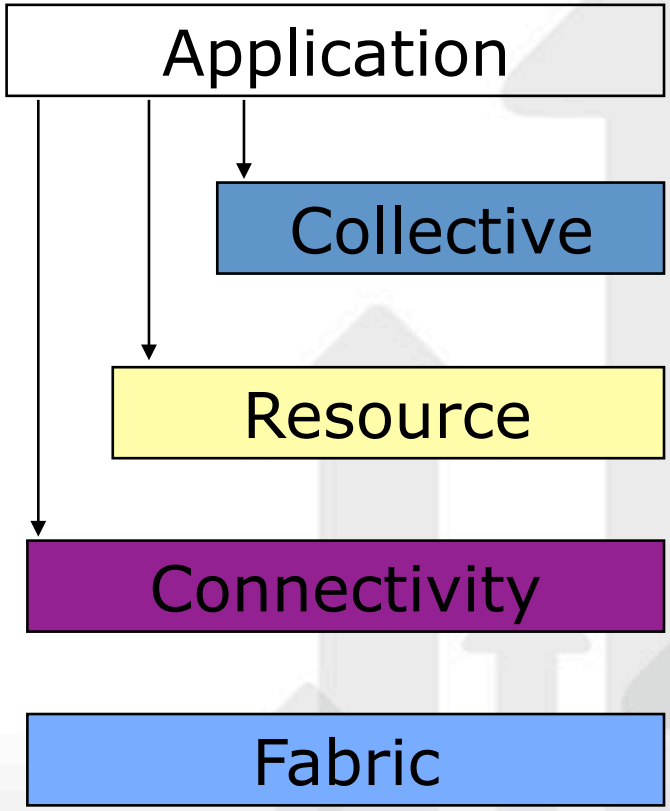
Ian Foster

Grid goals

Accelerate scientific discovery and innovation via

1) **On-demand computing**
   - Scale computing with need
   - Access to remote software

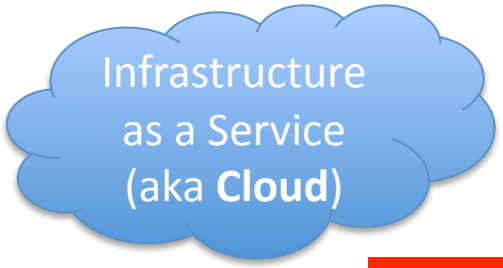
2) **Resource federation**
   - Computers
   - Data
   - Software
   - Instrumentation
   - People



Grid goals revisited: 2011

Accelerate scientific discovery and innovation via

1) **On-demand computing**
   - Scale computing with need
   - Access to remote software

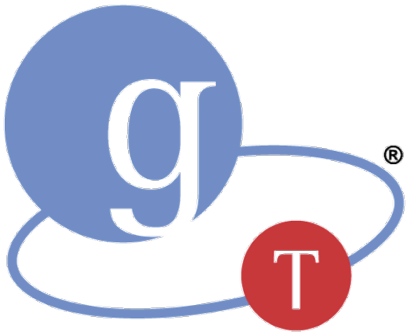


2) **Resource federation**
   - Computers
   - Data
   - Software
   - Instrumentation
   - People

Cloud = hosting
Grid = federation

Globus Toolkit

Build the Grid



Components for building custom grid solutions

globustoolkit.org

Globus Online

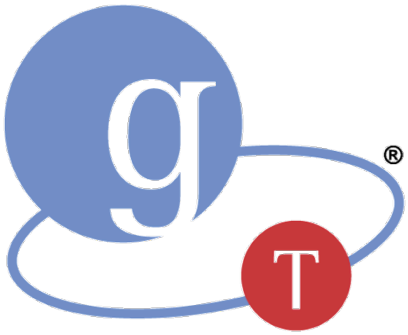
Use the Grid

Cloud-hosted file transfer service

globusonline.org

Globus Toolkit

Build the Grid



Components for building custom grid solutions

globustoolkit.org

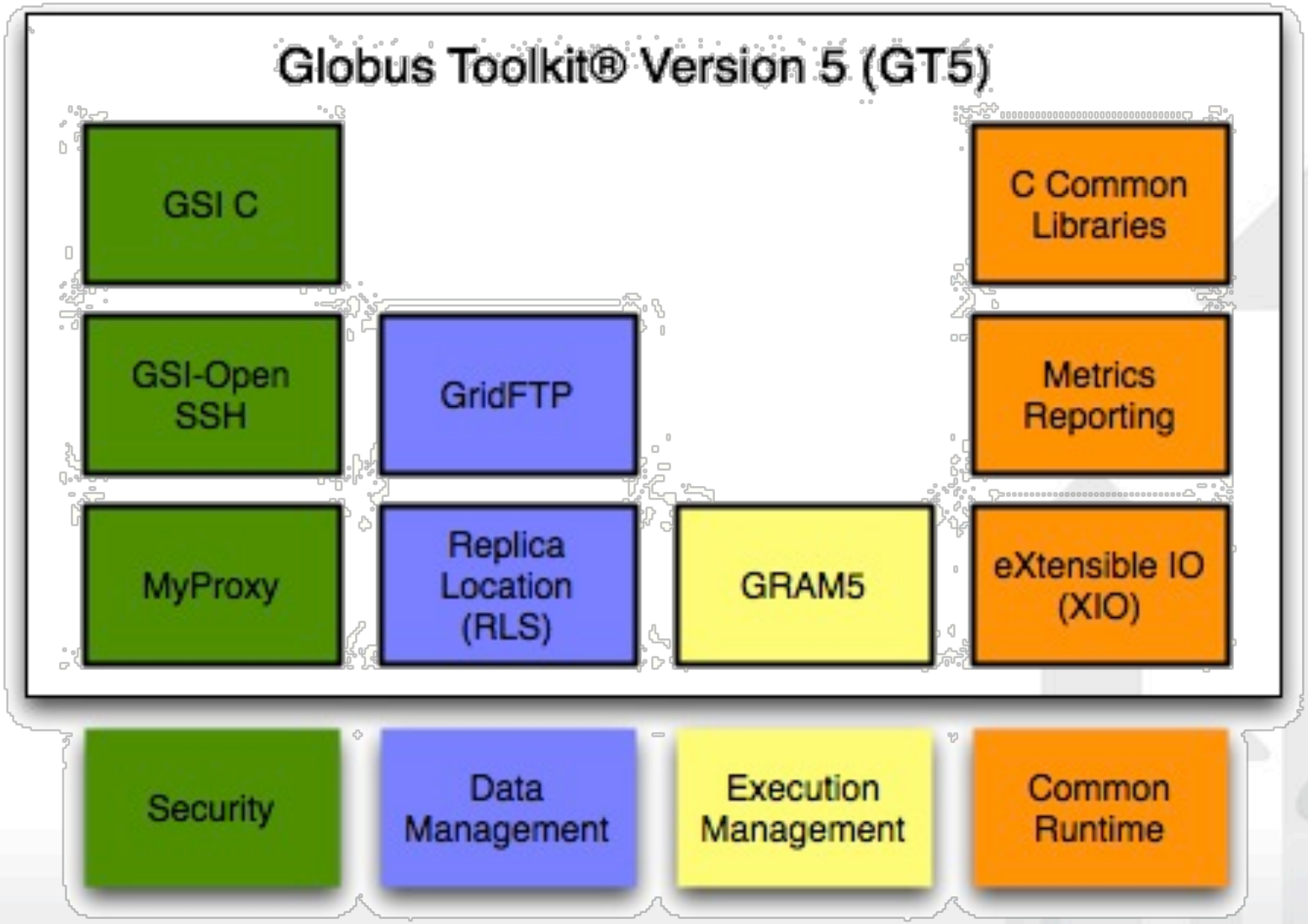
Globus Online

Use the Grid

Cloud-hosted file transfer service

globusonline.org

Globus Toolkit architecture overview



Other Globus Toolkit 4 components

- **jGlobus**
  - It's still there, just not in the figure!
  - GSI, Java clients for GRAM, GridFTP and MyProxy
- **Java Web Services core**
  - Limited adoption of WSRF
  - We recommend authoring Web Services with JAX-RS or JAX-WS
  - Crux project to support stateful Web Services not yet funded
- **WS-GRAM4**
  - We recommend migrating to GRAM5: better performance, reliability, and functionality
- **MDS monitoring and discovery service**
  - Integrated Information System (IIS) is an evolution of this

Recent Globus Toolkit developments

- **GridFTP**
  - Native packaging (in 5.2 alpha)
  - Globus Connect one-click install, one-paste config (later)
  - Data Channel Security Context support
  - Chrooting GridFTP server
    - Restrict access to a specific path
- **GRAM5**
  - GRAM2 improved and modernized (no Web Services inside)
  - Greater than 10x scalability than GRAM2 and roughly 10x reduction in resource consumption on the service host
  - Easier to debug (from working with TG)
    - Added RSL attribute to “save_job_description”

Coming soon

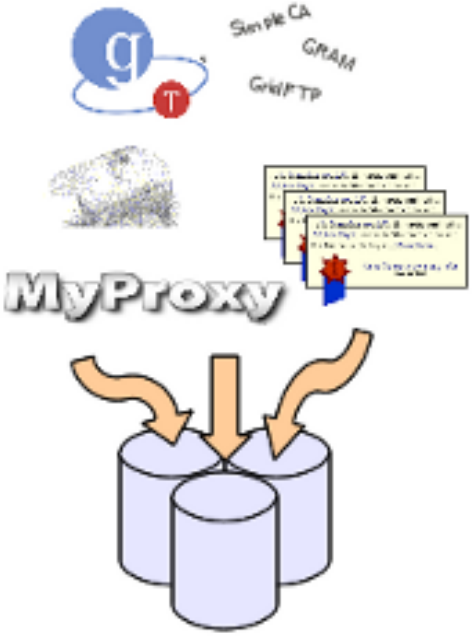
- **GRAM5**
  - Native packaging: RPM, Debian
  - Reduce complexity for sites updating GRAM
    - Modify LRM adapter interface to use callouts instead of patching source directly
  - Easier to debug (from working with TG) coming in 5.0.4
    - Added RSL attributes to control log location and level per job
  - Basic Execution Service (BES) interface – from IGE
  - Job Submission Description Language (JSDL) support – from IGE
- **jGlobus 2**
  - Security, GRAM and GridFTP interfaces
  - jGlobus 2 beta available, updates to latest SSL

DemoGrid (Borja Sotomayor)

```
[general]
organizations: a b
myproxy: yes
griduay: yes

[organization-a]
grid-users: 2
grid-users-auth: certs
gridftp: yes
lrm: condor
cluster-nodes: 2

...
```

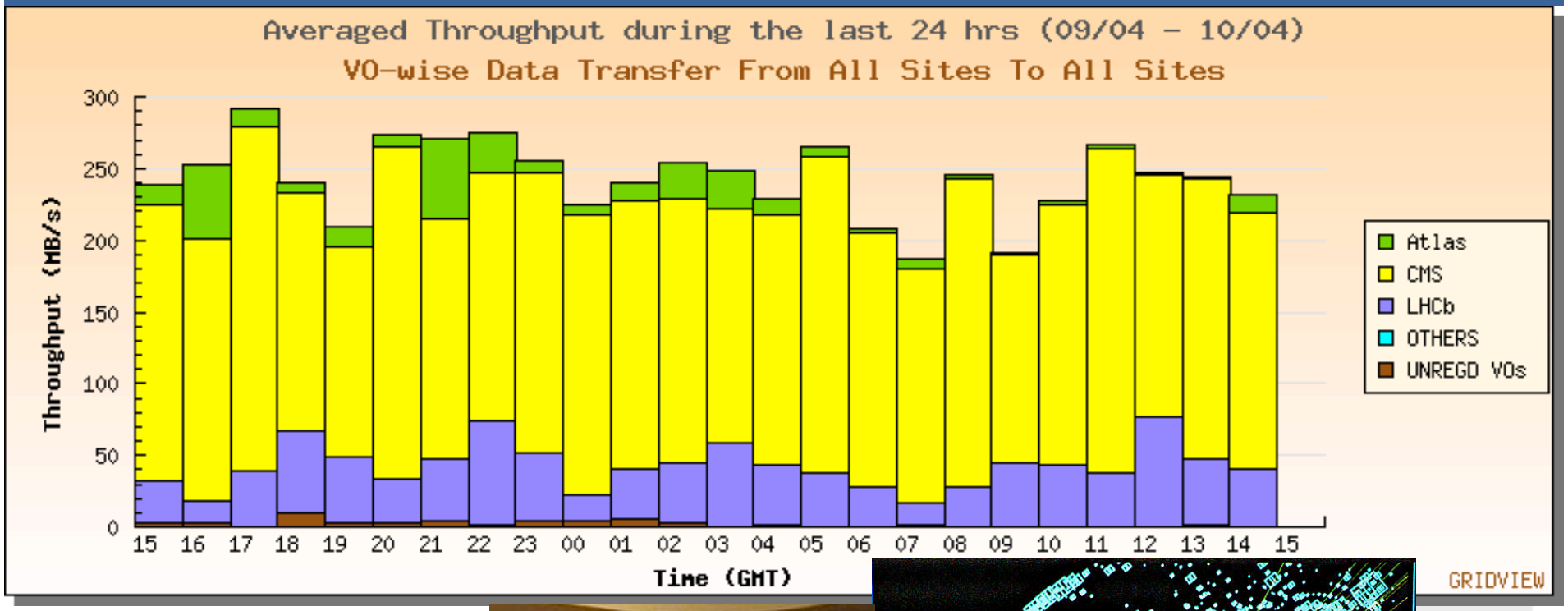


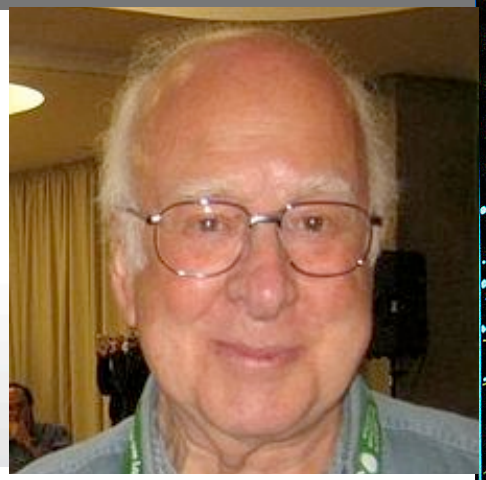
1 Specify a grid

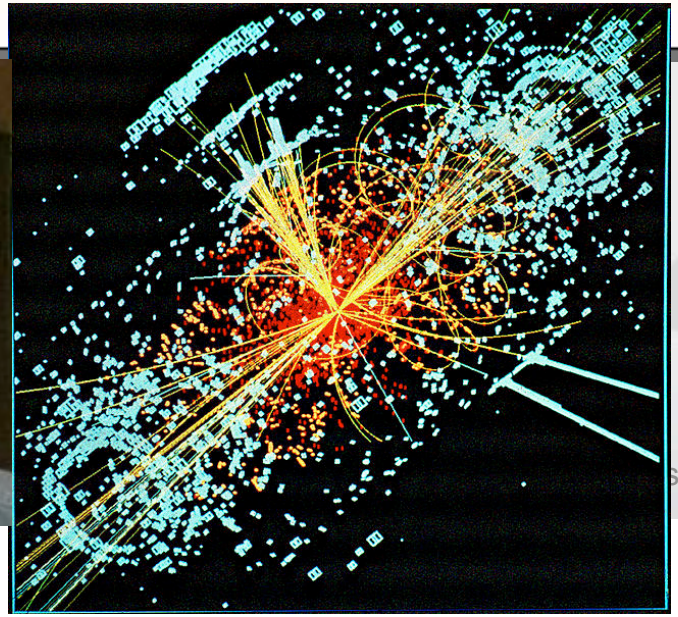
2 DemoGrid prepares the virtual machine images

3 Deploy your grid using EC2, Vagrant, or KVM

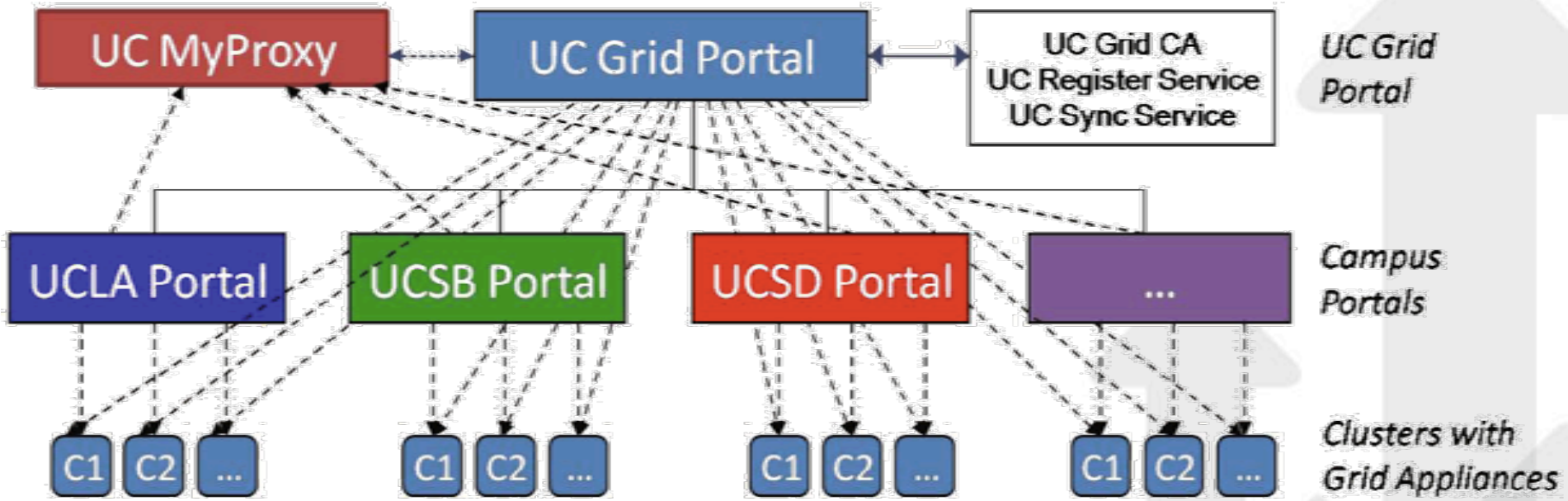
LHC Computing Grid







University of California (UC) Grid



The Nimbus Project

High-quality, extensible, customizable, open source implementation

Nimbus-Platform

Context Broker | Elastic Scaling Tools | Multi-Cloud Tools

Enable users to use IaaS clouds

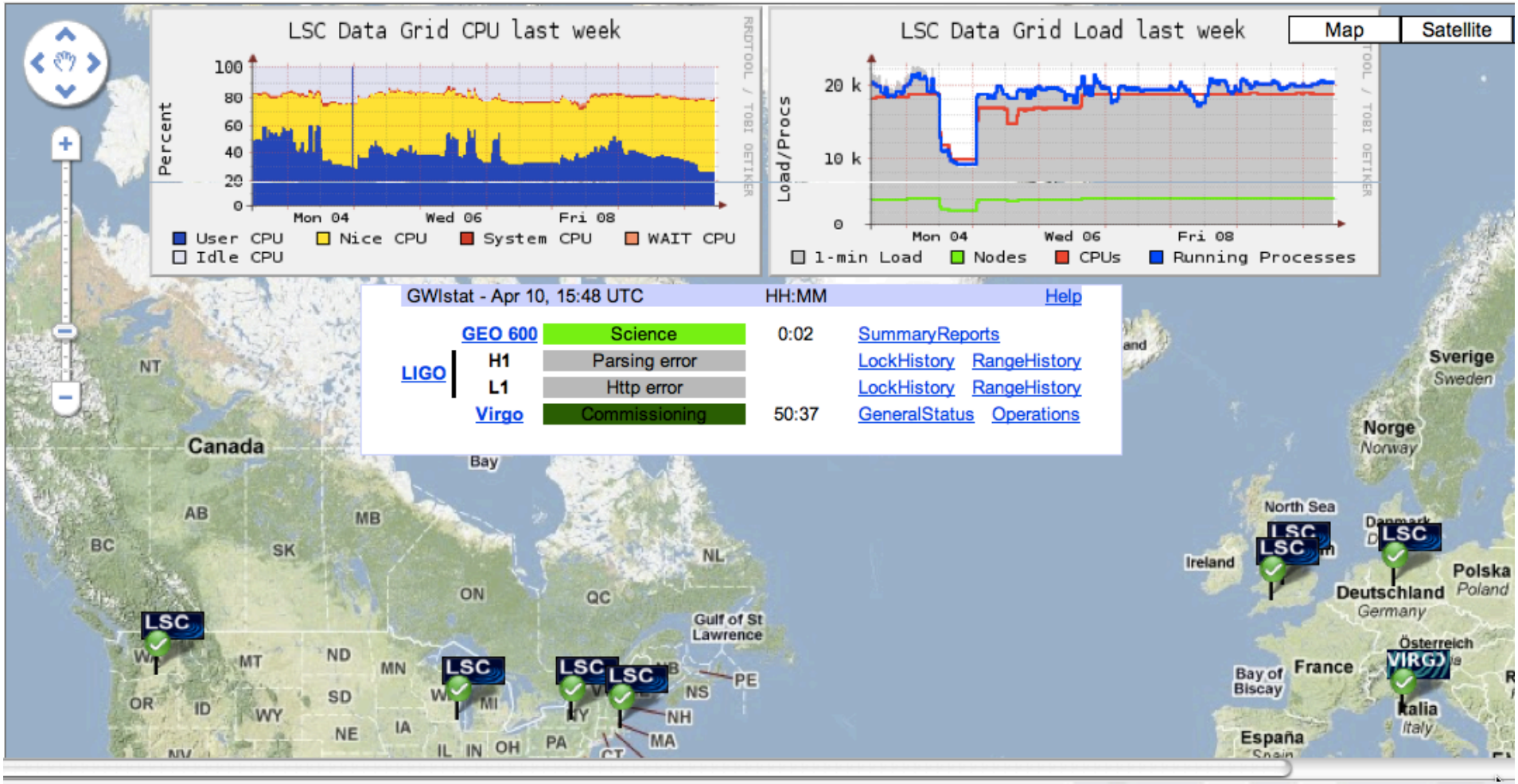
Nimbus-Infrastructure

Workspace Service | Cumulus

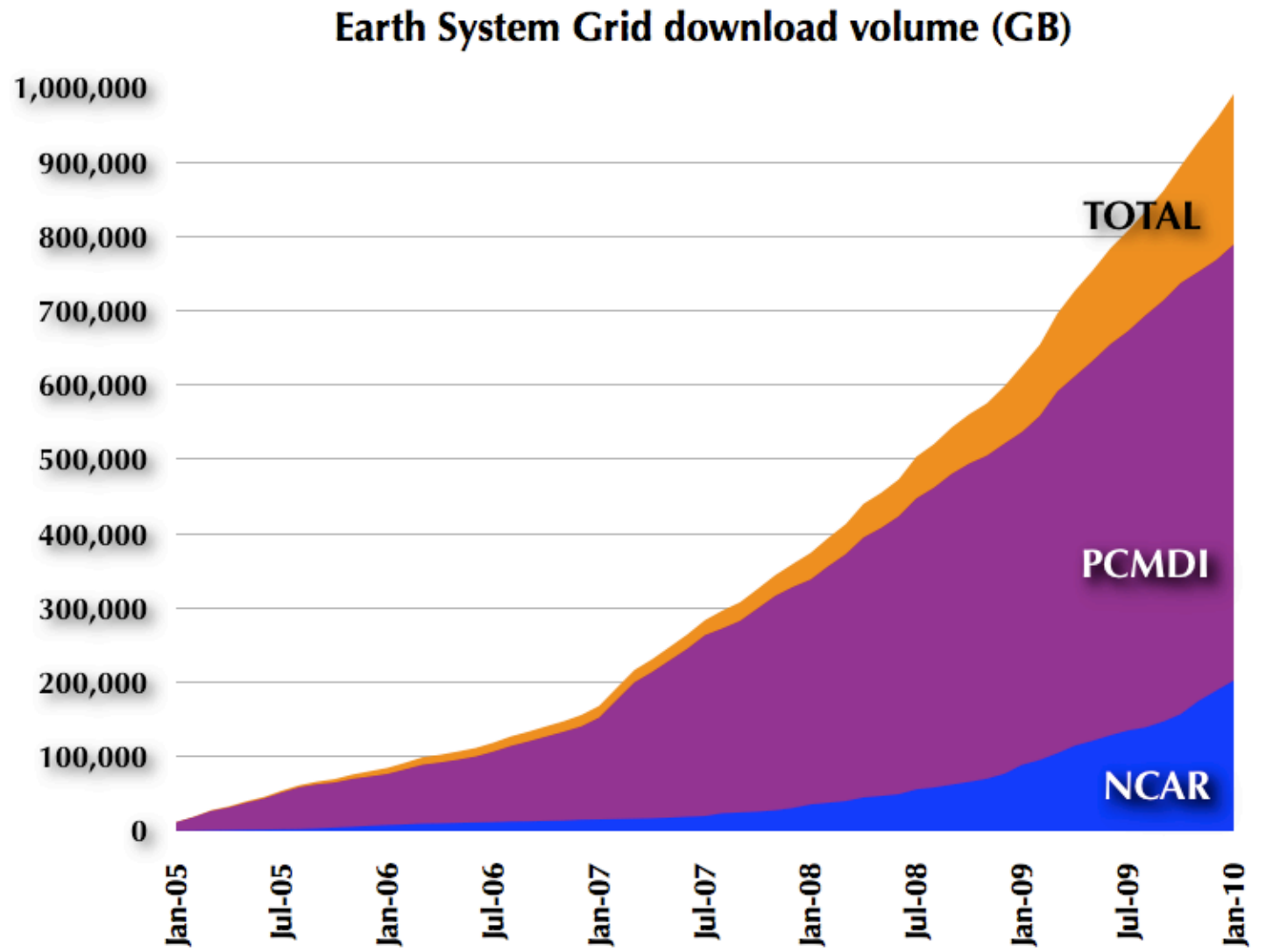
Enable providers to build IaaS clouds

Enable us to experiment with and evaluate cloud computing

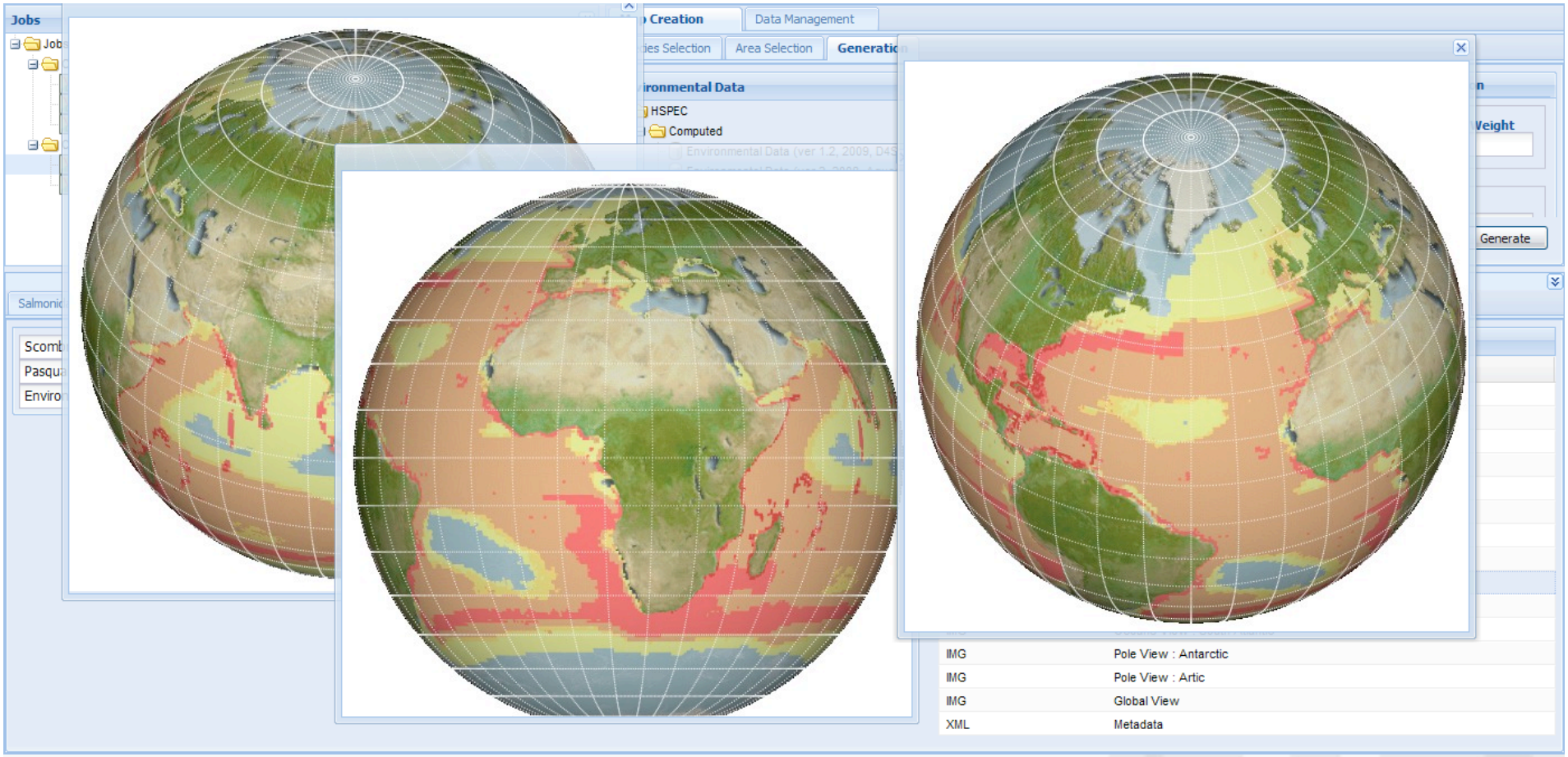
LIGO Scientific Collaboration



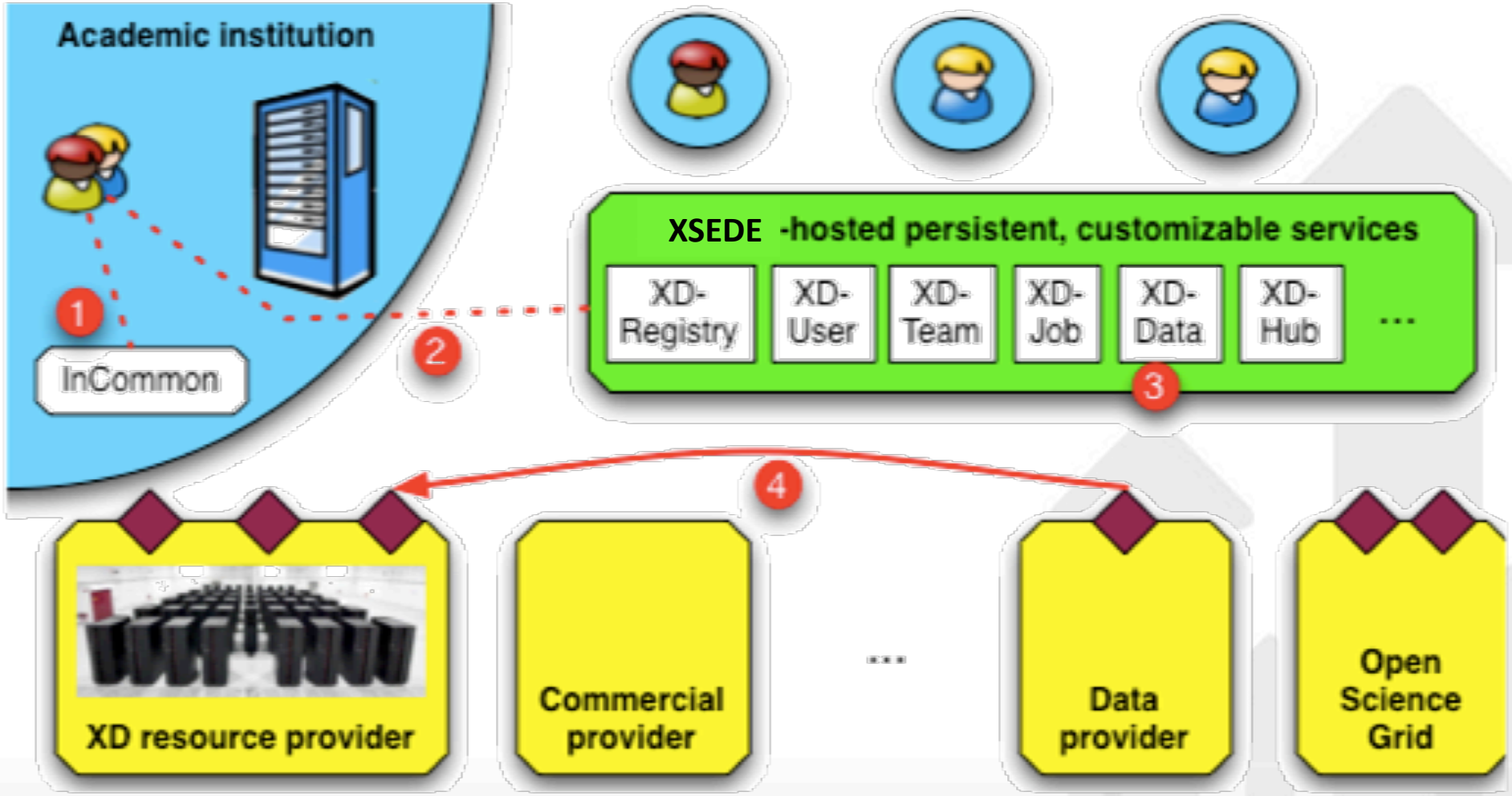
ESG has delivered a *Petabyte* of climate data to >20,000 users (Feb '10)



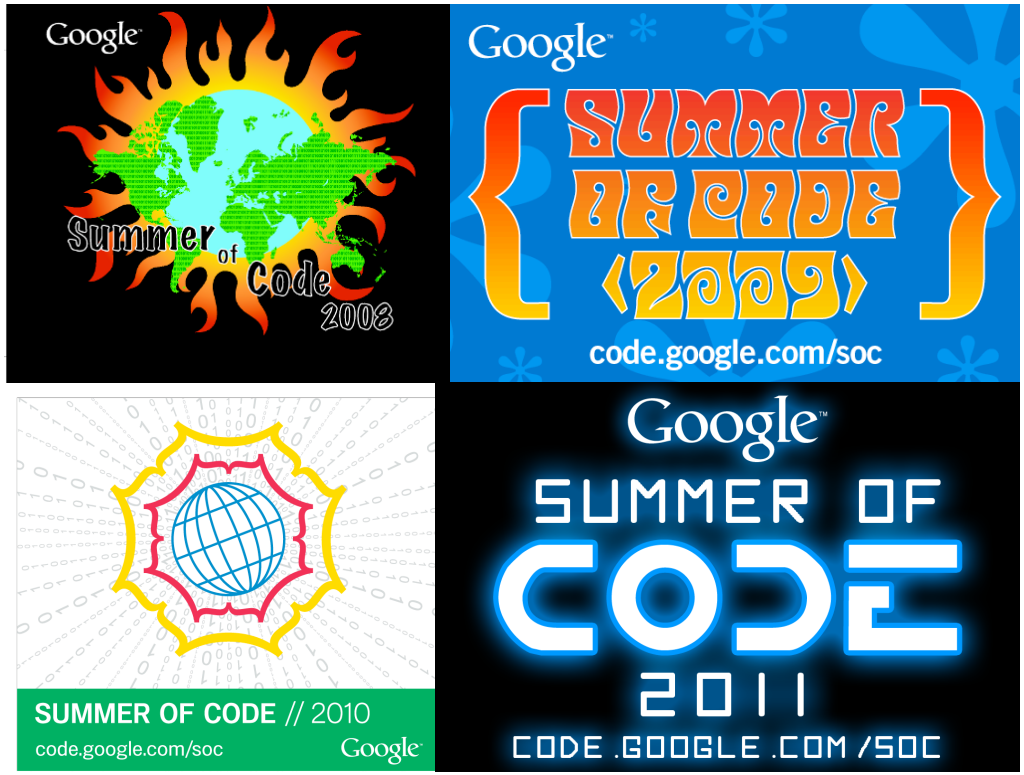
EU gCube project AquaMaps



NSF XD architecture incorporates GT, GO



Google Summer of Code



- Globus has been a mentoring organization in Google Summer of Code 2008, 2009, and 2010
- 29 students funded by Google to work on Globus technologies over the summer
- Recently accepted to GSoC 2011

NEWT - NERSC Web Toolkit

- NEWT - Web Service that makes NERSC HPC resources available as http URLs
- Build web applications through REST API
- User interacts with a web application that exposes the necessary components of the underlying application

  - Upload/download files
  - Authentication
  - Submit jobs
  - Accounting information
  - View Batch Queue
  - Key Value Store

- http://newt.nersc.gov

Built on top of Globus Tookit (GT5). Globus provides the underlying security, job and file transfer layers, without directly exposing this to end user

U.S. DEPARTMENT OF ENERGY | Office of Science

newt

BERKELEY LAB Lawrence Berkeley National Laboratory

Globus Toolkit

Build the Grid

Components for building custom grid solutions

globustoolkit.org

Globus Online

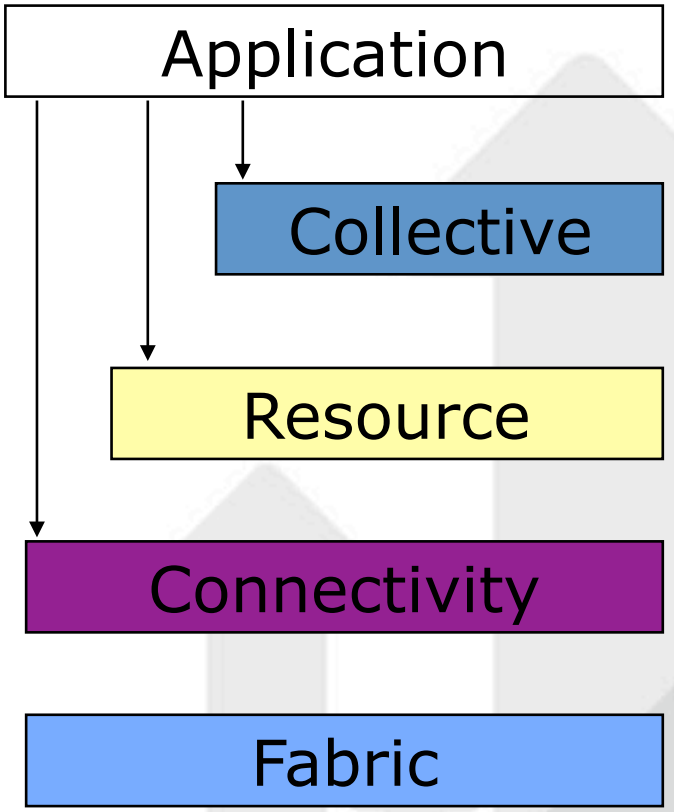
Use the Grid

Cloud-hosted
file transfer service

globusonline.org

Thinking about “small and medium labs”

- Big projects like LHC, LIGO, ESG, etc., can run **resource-level** services reliably—and build and operate effective **collective** services
- Small labs and collaborations have problems with both
- They need **solutions**, not toolkits —ideally **outsourced solutions**



Can we harness the power of the cloud to scale access to the grid?

Deliver research automation functions to small/medium labs via SaaS

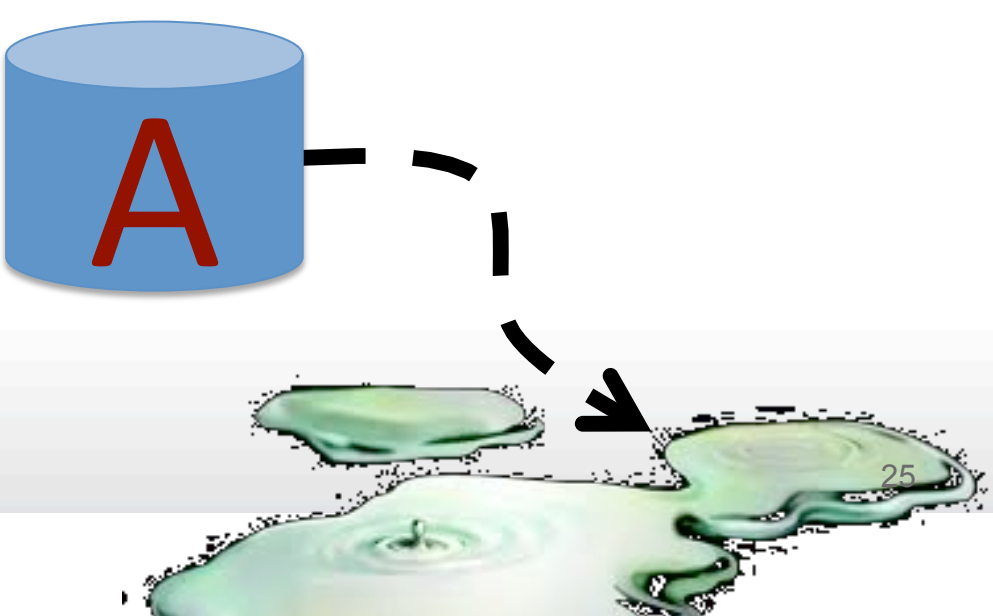
- **SaaS means:**
  1) Application own, delivered, managed by provider
  2) Single code base supports many users at once
  3) Application is properly Web architected
- **Has many potential advantages**
  - Leverage Web 2.0 to achieve extreme ease of use
  - Substantial economies of scale
  - Expert operations and support
  - Rapid software update
- **As well as challenges**
  - Paying for it. Security and privacy.

Time-consuming tasks in science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature
- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
- …

Starting with data movement

Discover endpoints, determine available protocols, negotiate firewalls, configure software, manage space, determine required credentials, configure protocols, detect and respond to failures, determine expected performance, determine actual performance, identify diagnose and correct network misconfigurations, integrate with file systems, …



Globus Online highlights

Command line interface
ls alcf#dtn:~
scp alcf#dtn:~/myfile \
nersc#dtn:~/myfile

HTTP REST interface
POST https://transfer.api.globusonline.org/ v0.10/transfer <transfer-doc>

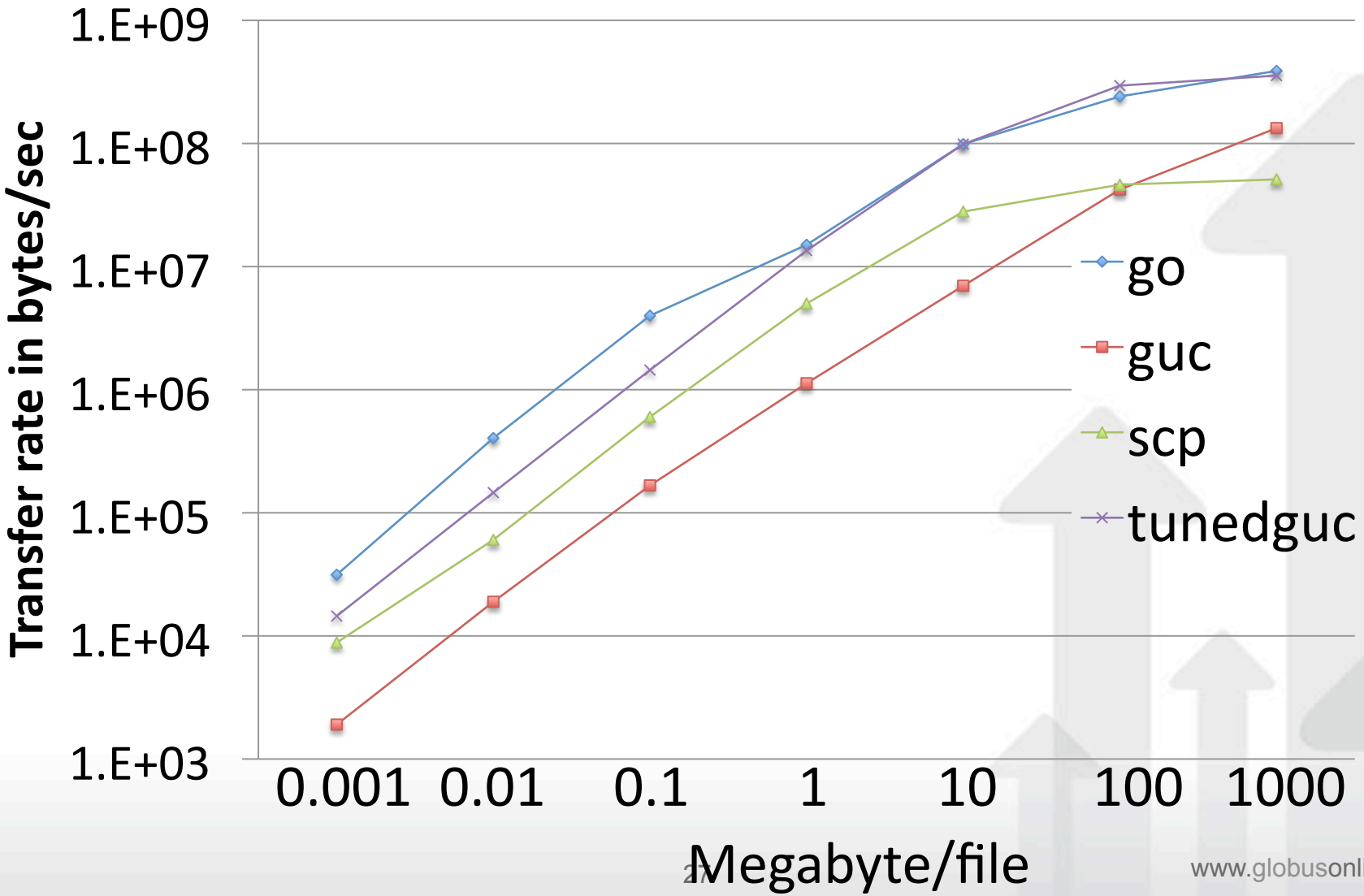
Fire-and-forget data movement
Many files and lots of data
Third-party transfers
Performance optimization
Expert operations and monitoring

GridFTP servers
FTP servers

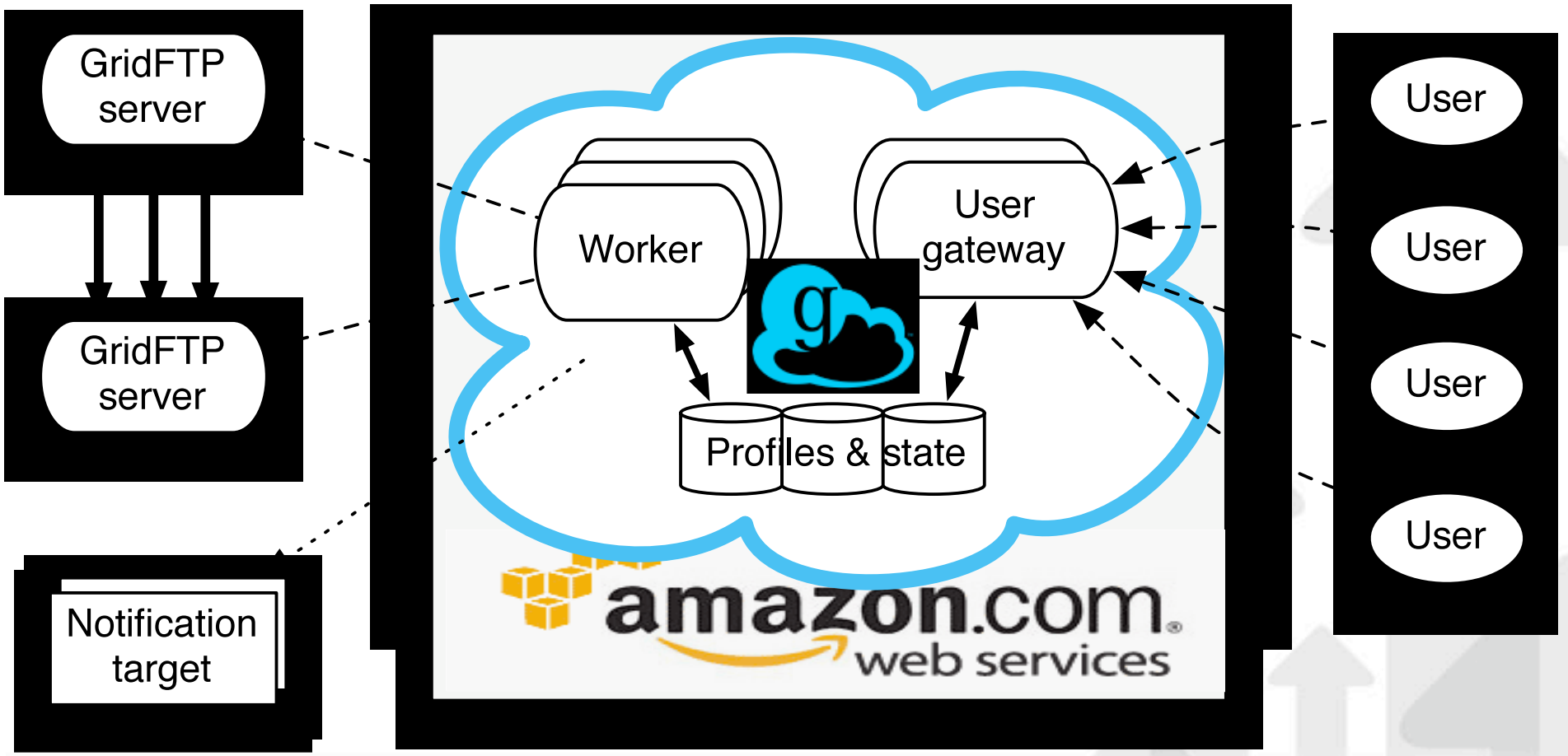
High-performance
data transfer nodes

Globus Connect
on local computers

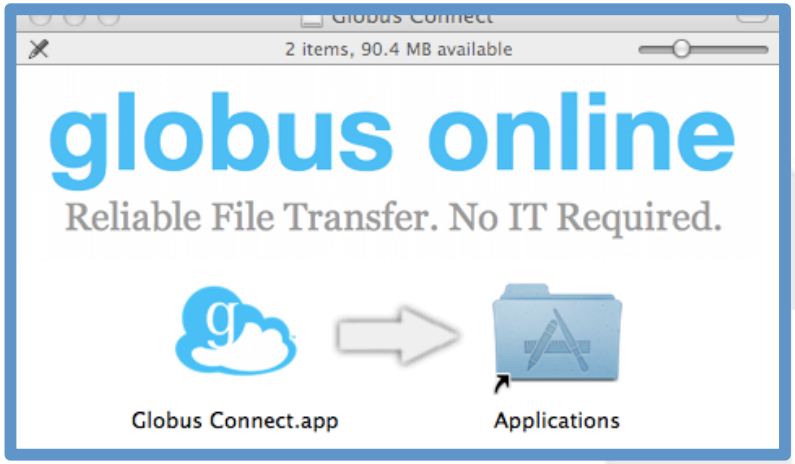
Globus Online is almost always faster than (even hand-tuned) globus-url-copy

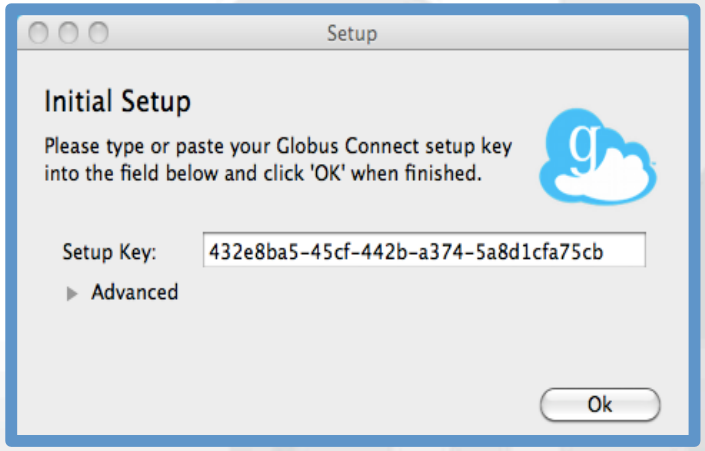


Globus Online architecture

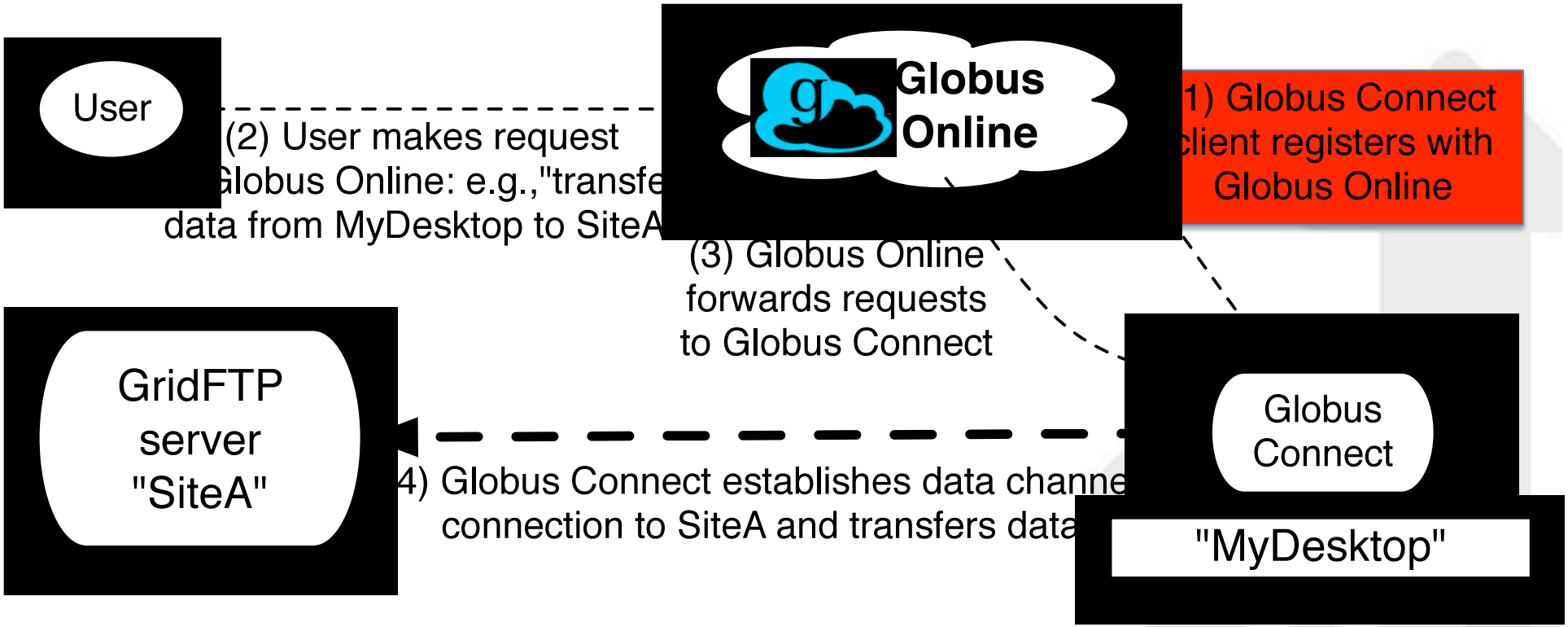


Globus Connect easy install

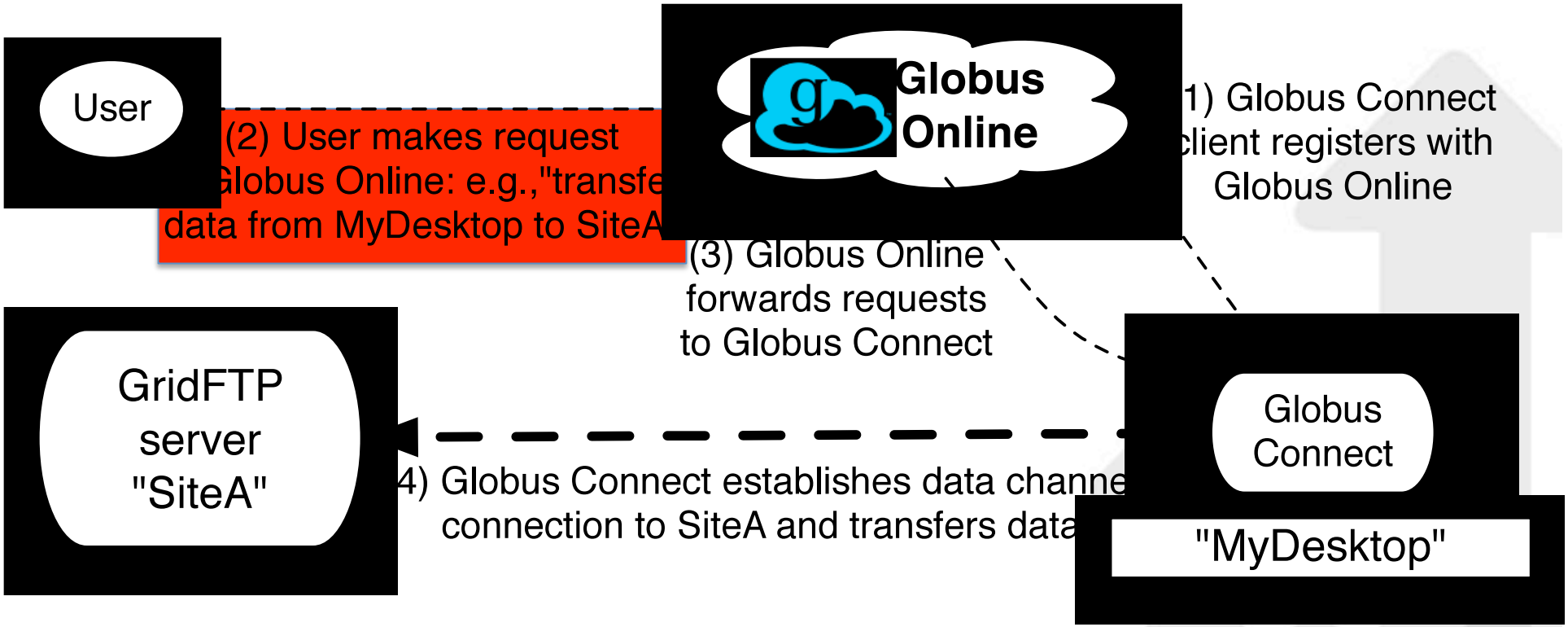




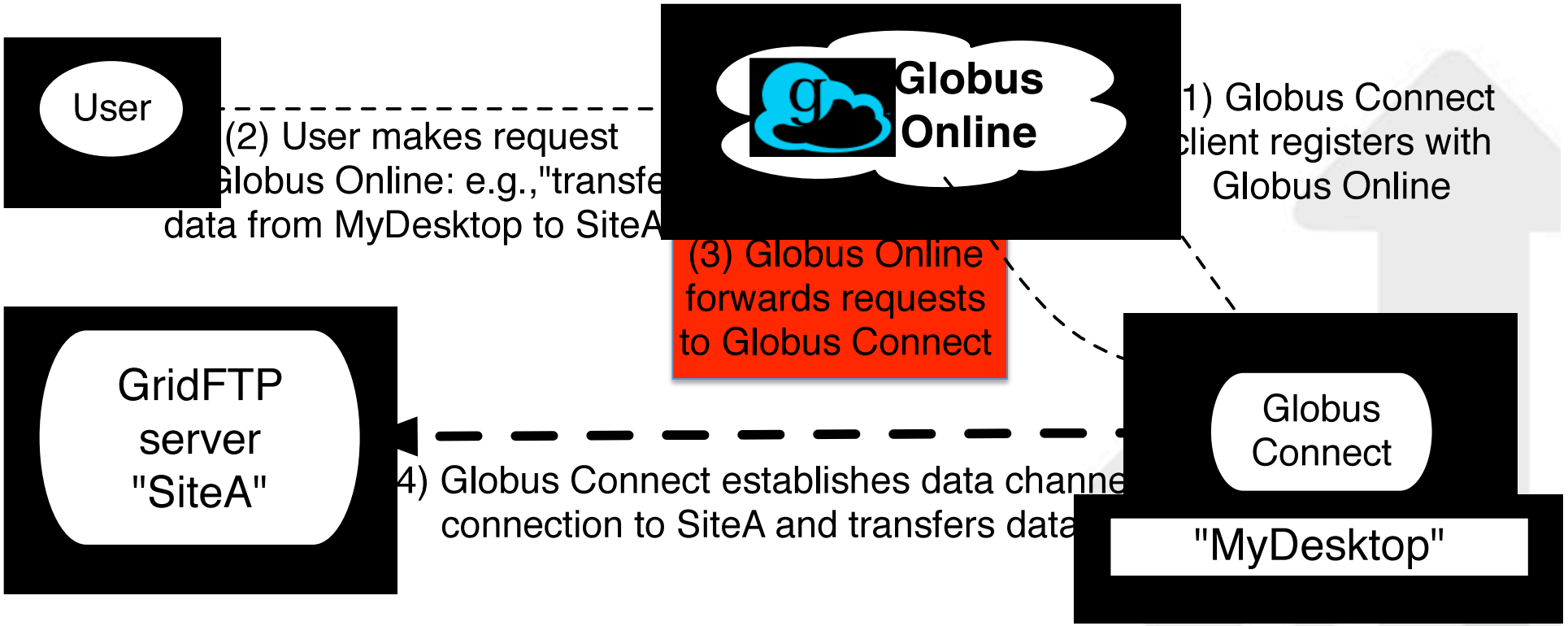
Globus Connect in action



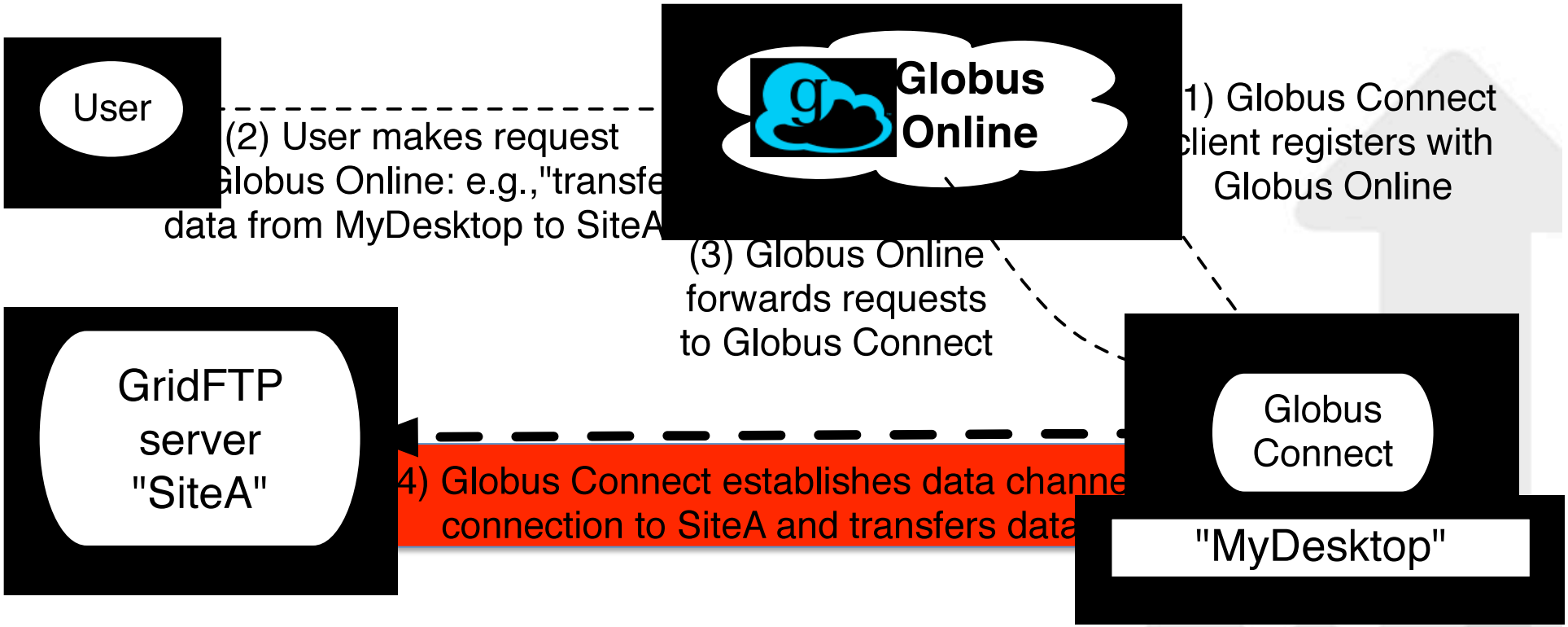
Globus Connect in action

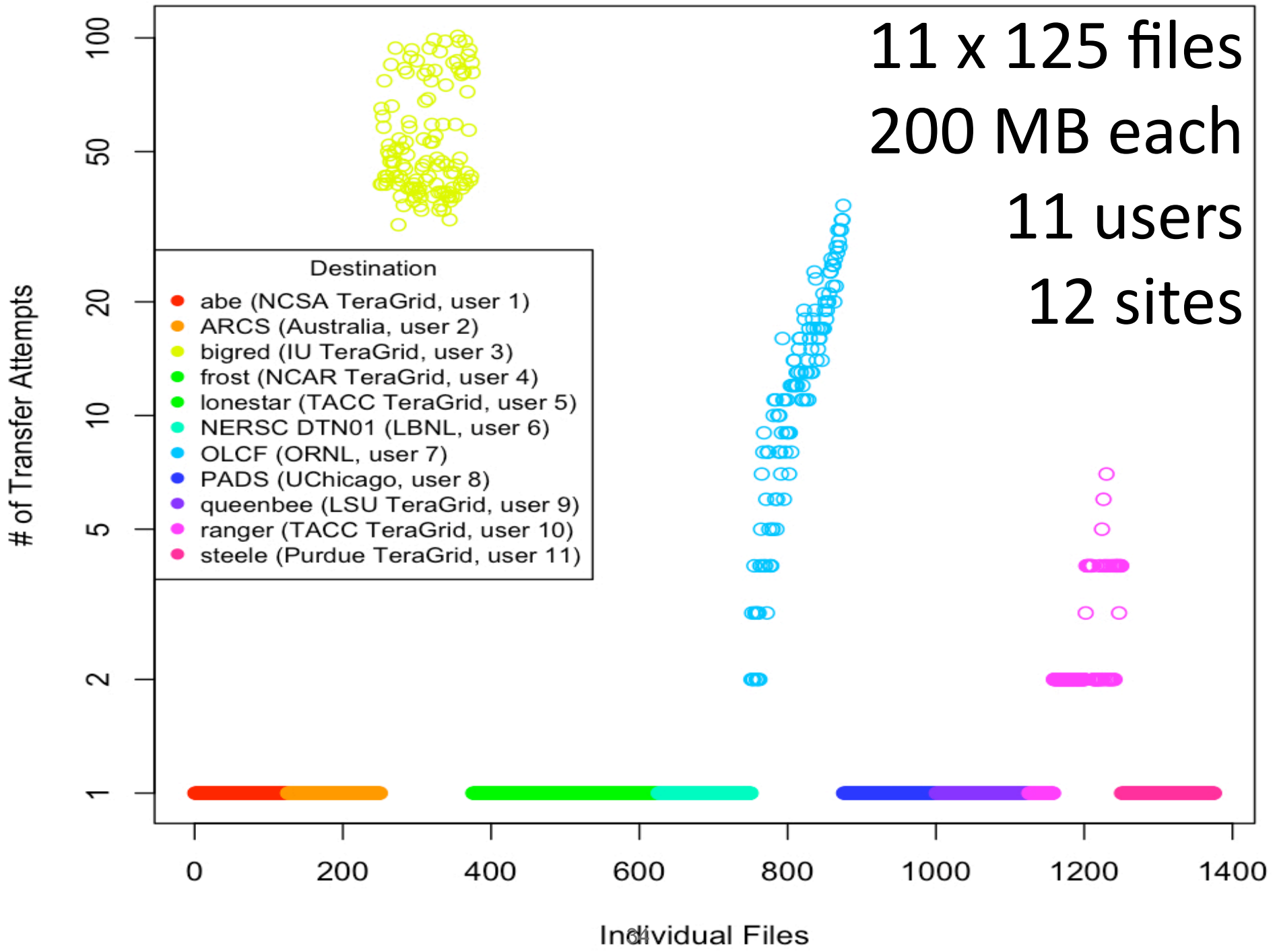


Globus Connect in action



Globus Connect in action




11 x 125 files
200 MB each
11 users
12 sites
Destination
abe (NCSA TeraGrid, user 1)
ARCS (Australia, user 2)
bigred (IU TeraGrid, user 3)
frost (NCAR TeraGrid, user 4)
lonestar (TACC TeraGrid, user 5)
NERSC DTN01 (LBNL, user 6)
OLCF (ORNL, user 7)
PADS (UChicago, user 8)
queenbee (LSU TeraGrid, user 9)
ranger (TACC TeraGrid, user 10)
steele (Purdue TeraGrid, user 11)
# of Transfer Attempts
1
2
5
10
20
50
100
0
200
400
600
800
1000
1200
1400
Individual Files

Time-consuming tasks in science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature
- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
- …

Time-consuming tasks in science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature
- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
- …

Looking to the future

Our goal: Accelerate discovery and innovation by providing research IT as a service

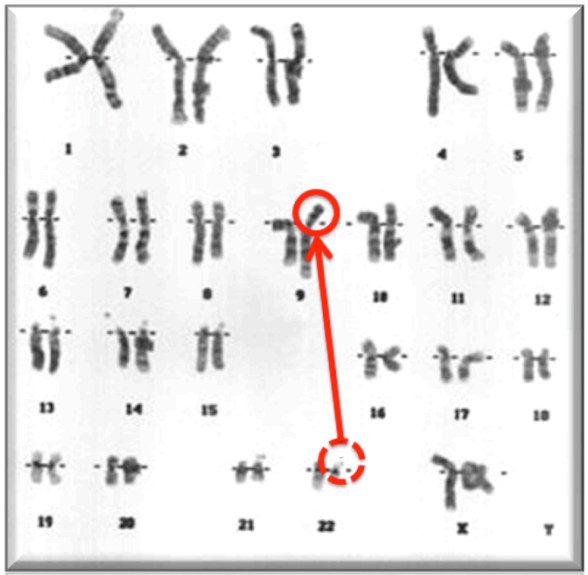
> *"Civilization advances by extending the number of important operations which we can perform without thinking of them"*
>
> Alfred North Whitehead , 1911

The opportunity

To leverage software-as-a-service (SaaS) to greatly accelerate the pace of discovery and innovation worldwide, by

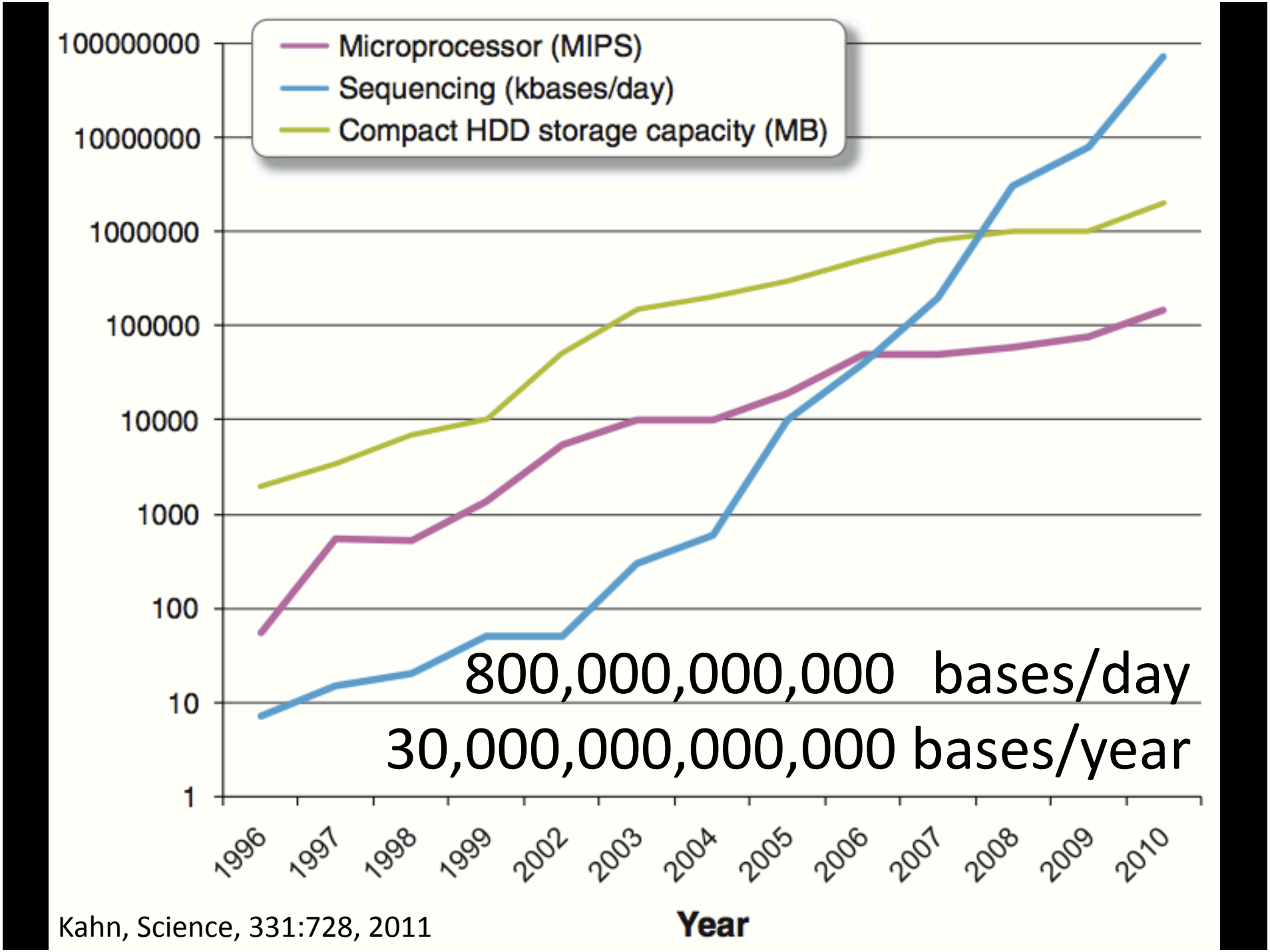
- providing millions of researchers with unprecedented access to powerful research tools, and
- enabling a massive shortening of cycle times in time-consuming research processes

Pattern recognition leads to discovery and cures



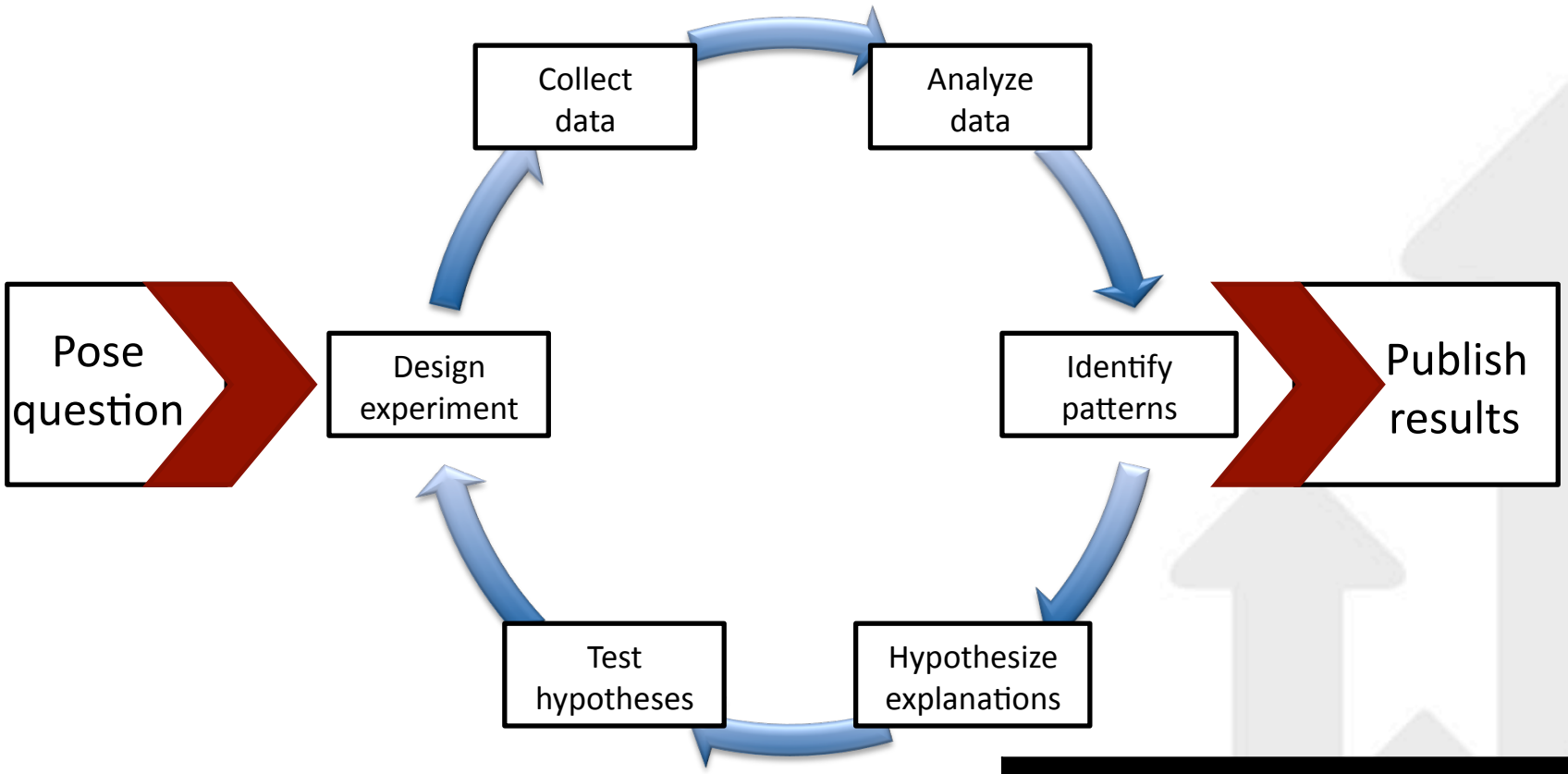
In 1972, Janet Rowley discovered that chromosome abnormalities can cause cancer

Many other breakthroughs are urgently needed



Kahn, Science, 331:728, 2011

Research is an iterative, time-consuming process



Massive worldwide opportunity

10+ million researchers
Millions of research labs

Technology evolution creates new opportunities

New opportunities for discovery and innovation		GO enables researchers to seize opportunities
•Massive deluge of data from many new instruments •Exponential increase in computing power •Pervasive collaboration among distributed teams	+	•Creation and management of massive data collections •Data analysis, mining, and simulation capabilities •Selective sharing and secure information exchange

=

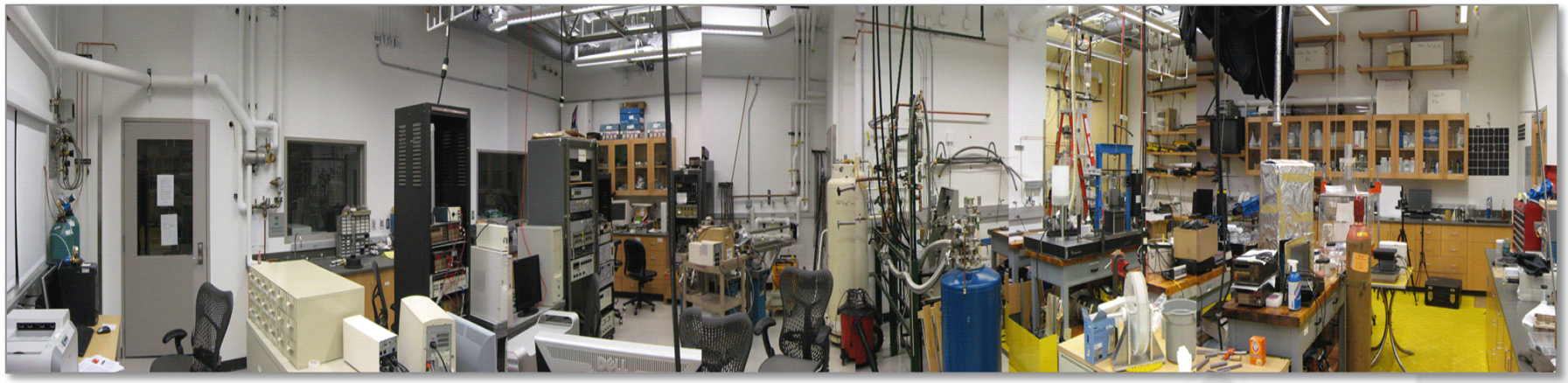
Ability to rapidly identify, analyze, and validate patterns accelerates discovery and innovation

Researchers lack IT support infrastructure

- **Almost all research performed in small laboratories**
- **Researchers are trained in their field, not in IT**
  - They are not experts in collecting, moving, storing, indexing, analyzing, mining, sharing, updating, publishing, and archiving massive amounts of data
- **Only limited capital is available for them to spend on data and IT support**
- **Investment is spent on traditional research tools (e.g., microscopes)—but the world is changing**
  - Now need substantial and sophisticated IT to perform research, data manipulation, data mining, collaboration

Researchers lack IT support infrastructure

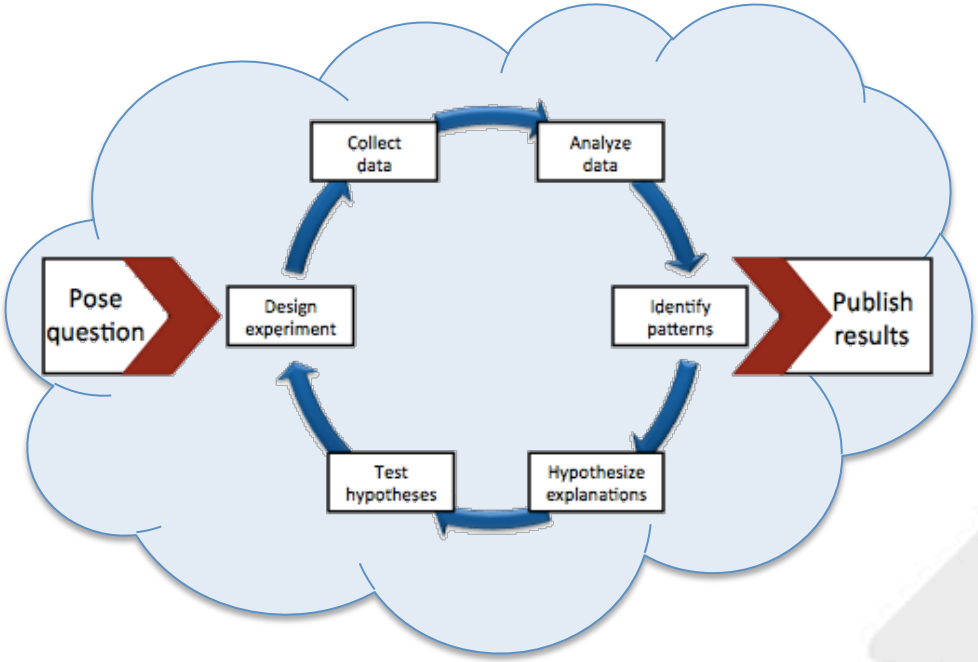
Traditional lab investment



Area of need

Lab data processing infrastructure

We believe the solution is to leverage SaaS to deliver missing IT



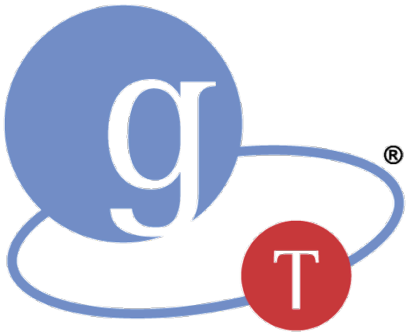
- **Remove need for researchers to create and manage their own IT infrastructure**
- **Deliver capabilities not achievable in individual labs**
- **SaaS model that is cost effective on a pay-as-you-go basis**

Summary

- **Data and technology proliferation creates huge opportunities for new discoveries and innovations**
- **But researchers lack the IT skills, tools, and resources ($) to leverage these opportunities**
- **We propose to solve this problem by providing missing IT to researchers via a cost-effective SaaS platform**
- **This new approach to research IT can greatly accelerate discovery and innovation worldwide, by**
  - providing millions of researchers with unprecedented access to powerful research tools, and
  - enabling a massive shortening of cycle times in time-consuming research processes

Globus Toolkit

Build the Grid



Components for building custom grid solutions

globustoolkit.org

Globus Online

Use the Grid

Cloud-hosted file transfer service

globusonline.org

Program for the rest of Monday

3:00p – 4:00p	**Globus Toolkit Updates** *Led by: Stuart Martin, UChicago* • CILogon, GridShib, MyProxy – Jim Basney, NCSA • jGlobus – Mike Russell, UChicago • Native packaging – Stuart Martin, UChicago • GRAM – Stuart Martin, UChicago • IIS – JP Navarro, Argonne Abstract: This session will update participants on recent developments and current status of core Globus Toolkit components and GT 5.2 native packaging.
4:00p – 5:00p	**Globus Community Updates & User Experiences** *Led by: Stuart Martin, UChicago* • VDT – Alain Roy, University of Wisconsin-Madison • OGSA-DAI – Charaka Palansuriya, EPCC • IGE – Alexander Papaspyrou, TU Dortmund • caBIG – Ravi Madduri, Argonne Abstract: This session will update participants on the experiences of select Globus communities and users.
5:00p	Adjourn
6:30p – 9:30p (est.)	**GlobusWORLD hosted party**

Program for Tuesday morning

7:30a – 8:30a	Registration & Continental breakfast
8:30a – 10:00a	**Globus File Transfer Updates & User Experiences** *Led by: Steve Tuecke, UChicago* • Globus Online file transfer – Steve Tuecke, UChicago • ESG – Rachana Ananthakrishnan & Neill Miller, Argonne • GridFTP – Raj Kettimuthu, Argonne Abstract: This session will provide attendees with a better understanding of new and soon-to-be-released features, functionality and use of Globus Online file transfer service.
10:00a – 10:30a	Coffee Break
10:30a – 12:00p	**Globus Community Updates & User Experiences** *Led by: Paul Davé, UChicago* • ARCS – Graham Jenkins, VPAC • NERSC – Shreyas Cholia, LBL • iBi – Brigitte Raumann, UChicago • GARUDA – Prahlada Rao, C-DAC Bangalore Abstact: This session will update participants on the experiences of select Globus communities and users.

Program for Tuesday afternoon

1:30p – 3:00p	**Globus Online – New Features and Future Enhancements** *Led by: Steve Tuecke, UChicago* • Globus Online roadmap – Steve Tuecke, UChicago • BIRN – Carl Kesselman, USC ISI • Virtual Endpoints – Gopi Kandaswamy, USC/ISI • Grisu – Markus Binsteiner, Centre for eResearch, University of Auckland Abstract: This session will introduce and demonstrate new Globus Online features.
3:00p – 3:30p	Coffee Break
3:30p – 4:15p	**Community Technology Updates** *Led by: Vas Vasiliadis, UChicago* • Condor – Zachary Miller, University of Wisconsin–Madison • Glidein WMS – Parag Mhashilkar, Fermi • Data–flow Parallelism – Esma Yildirim, SUNY – Buffalo Abstract: This session will update participants on Globus technology extensions, enhancements, and integration.
4:15p – 5:00p	**Session: Recap and Q&A** *Led by: Ian Foster, Argonne and UChicago* Abstract: This session will briefly recap the conference and provide time for general questions and answers.

Tutorials on Wednesday

8:30a – 10:00a	**Globus Online Introduction** *Led by: Lisa Childers, Argonne* Abstract: This tutorial will provide an introduction to Globus Online (http://www.globusonline.org), the latest addition to the Globus software suite. The session will begin with context-setting material, including an overview of the motivation for Globus Online and key design concepts.
10:00a – 10:30a	Coffee Break
10:30 – 12:00p	**Enabling Your HPC Cluster with Globus** *Led by: Borja Sotomayor, UChicago* Abstract: This tutorial will demonstrate how to install and administer Globus Toolkit on a cluster, and how to configure GridFTP (and related security components) so that an HPC resource can be used with Globus Online.
12:00p – 1:30p	Lunch
1:30p – 3:00p	**Globus Online Advanced CLI and Scripting** *Led by: Lisa Childers, Argonne* Abstract: This tutorial will explore advanced usage of Globus Online via the Command Line Interface (CLI), and how this interface can be used for scripted usage of Globus Online.
3:00p – 3:30p	Coffee Break
3:30p – 5:00p	**Globus Online Transfer REST API** *Led by: Bryce Allen, UChicago* Abstract: This tutorial will teach attendees how to use the Globus Online Transfer REST API, for programmatic interaction with Globus Online. Examples will demonstrate using the Transfer REST API to integrate Globus Online with Java and Python clients and Web-based portals.

Other related meetings

- **Cloud Computing and Applications**
  - Wednesday
- **Workshop on High Performance Applications of Cloud and Grid Tools**
  - Thursday

Follow links from www.globusworld.org for details

Contests at GlobusWORLD

- **Data Challenge**
  - **What you do:** Move the most data during first two days of GlobusWORLD
  - **What you get:** iPad 2

- **Story Challenges**
  - **What you do:** Tell the best Globus Online user story…
    - Most Innovative Use
    - Most Ambitious Vision
  - **What you get:** iPod Touch

- **How to Participate**
  - Visit www.globusonline.org/gw11contests