Computation Institute

Big process for big data

Process automation for data-driven science

Ian Foster

Computation Institute

Mathematics and Computer Science Division

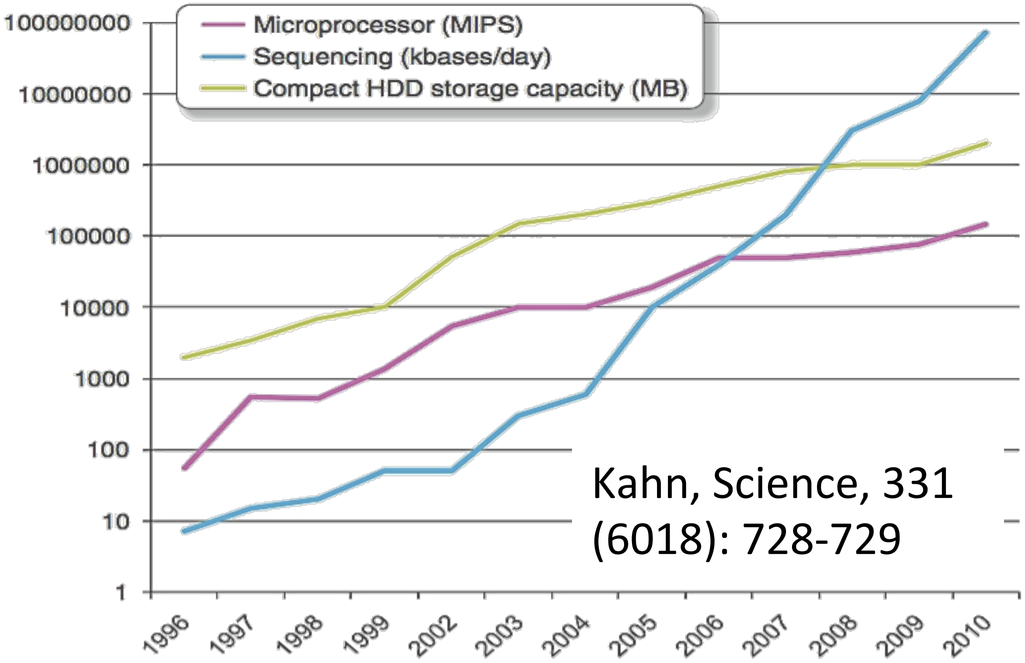
Department of Computer Science

Argonne National Laboratory & The University of Chicago

Talk at University of Joint Lab Workshop, Argonne, November 20, 2012

Argonne NATIONAL LABORATORY

A productivity crisis in research

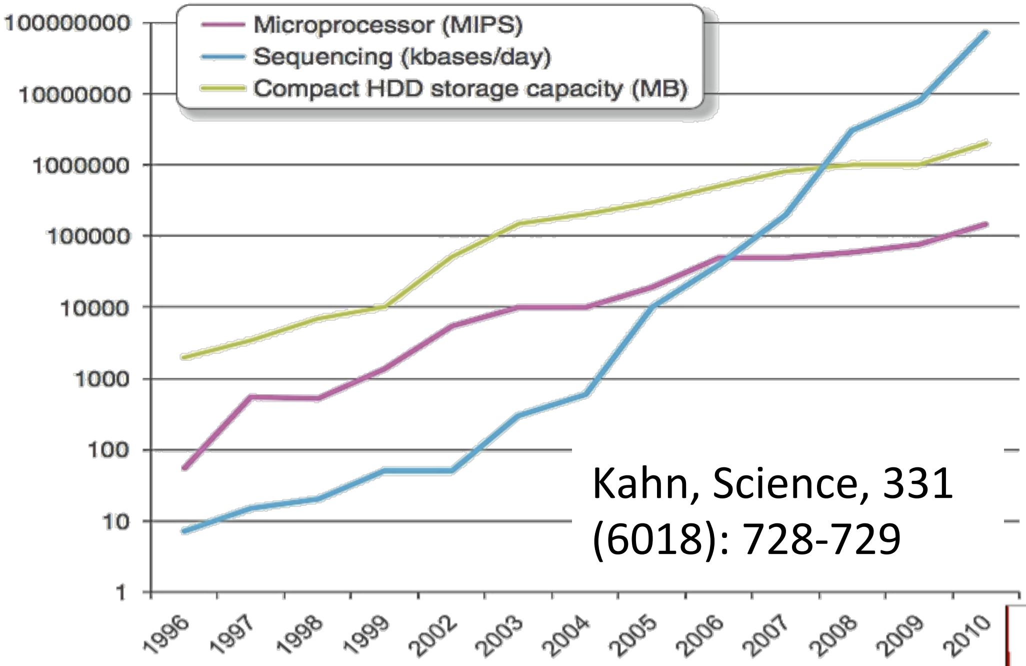


Kahn, Science, 331 (6018): 728-729

Data volumes are growing **much faster** than Moore’s law …

(10,000x more over last 6 years for genome data)

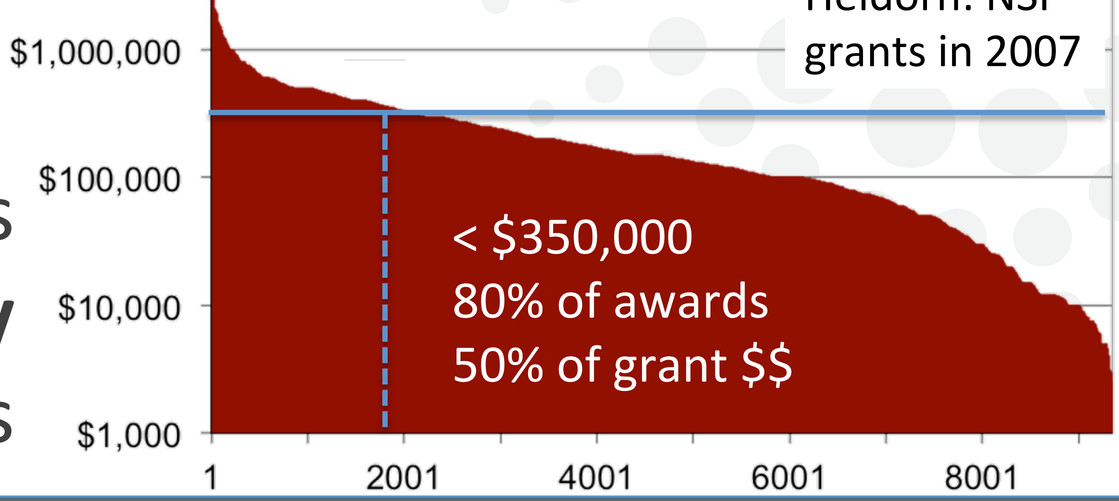
A productivity crisis in research



Kahn, Science, 331 (6018): 728-729

Data volumes are growing **much faster** than Moore's law …

(10,000x more over last 6 years for genome data)



Heidorn: NSF grants in 2007

< $350,000
80% of awards
50% of grant $$

But most labs have **extremely limited** resources

Automation and outsourcing are key

- **Automation** is required to apply more sophisticated methods to far more data

Automation and outsourcing are key

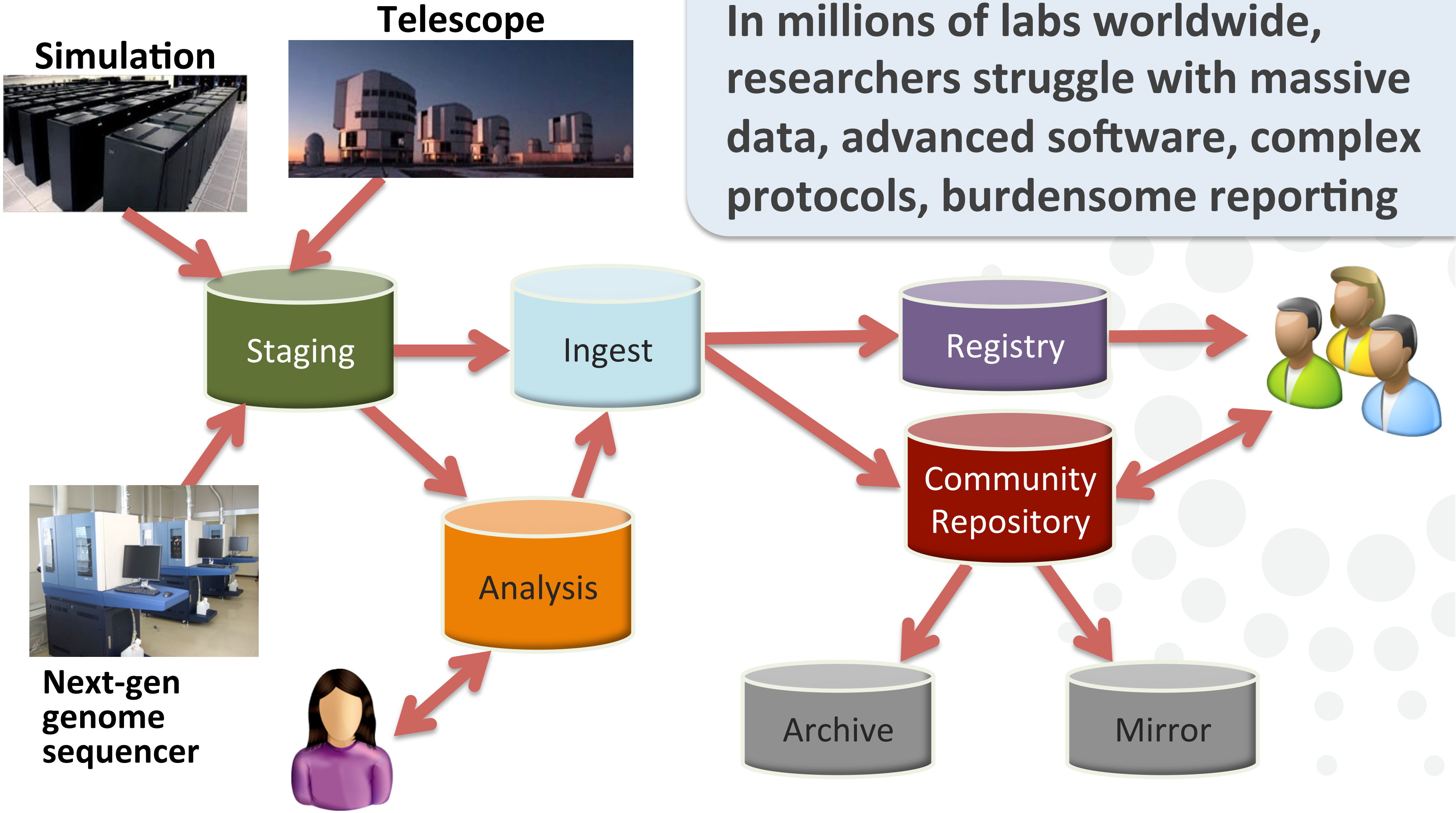
- **Automation** is required to apply more sophisticated methods to far more data

- **Outsourcing** is needed to achieve economies of scale in the use of automated methods

They work for small and medium business

The research data lifecycle

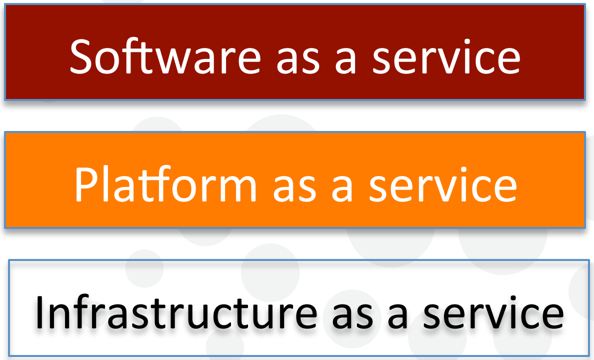
In millions of labs worldwide, researchers struggle with massive data, advanced software, complex protocols, burdensome reporting



Accelerate discovery and innovation by outsourcing difficult tasks

Research strategy

- Identify **time-consuming activity** that appears amenable to automation and outsourcing
- Implement activity as a high-quality, low-touch **SaaS solution** with high economies of scale
- Evaluate **usage and performance**
- Extract common elements as a **research automation platform**
- Repeat


Software as a service

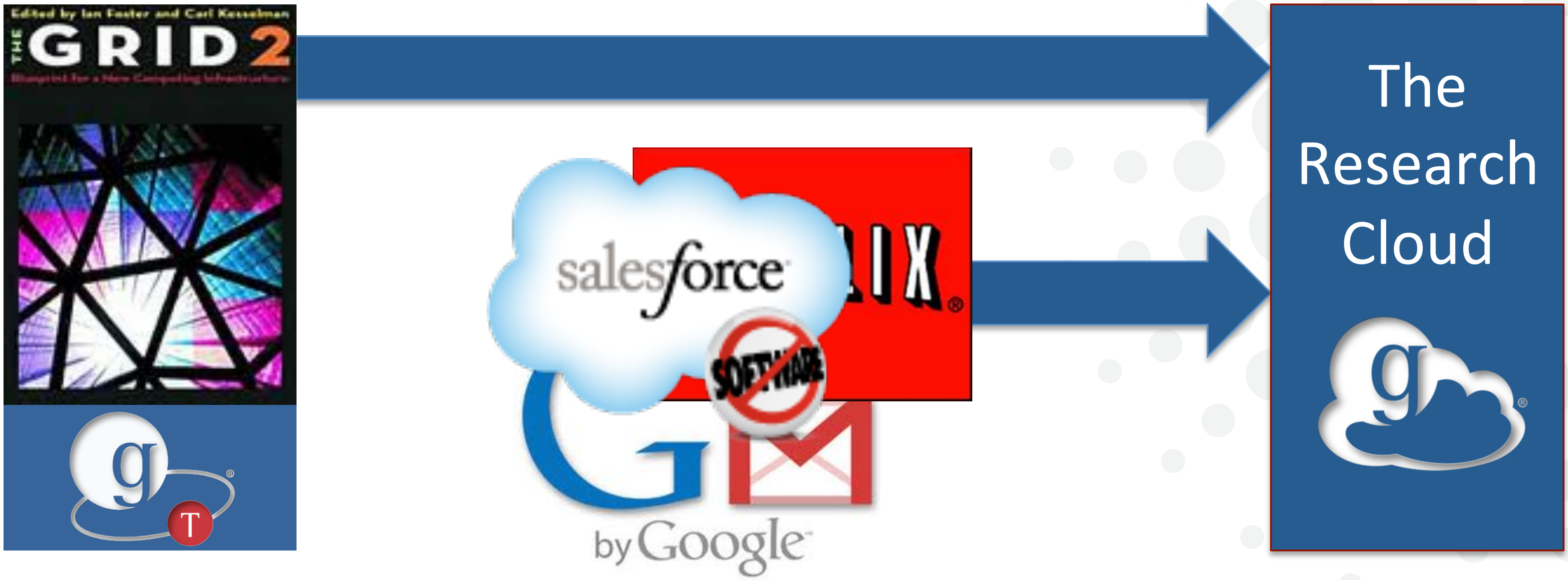
Platform as a service

Infrastructure as a service

Bonus question: Identify methods for delivering SaaS solutions in sustainable manner

Grid meets Cloud

Millions of researchers worldwide need advanced IT to tackle important and urgent problems


Edited by Ian Foster and Carl Kesselman
THE GRID 2
Blueprint for a New Computing Infrastructure
salesforce
LIX
SOFTWARE
by Google

The Research Cloud

Characterizing big process requirements

Simulation

Telescope

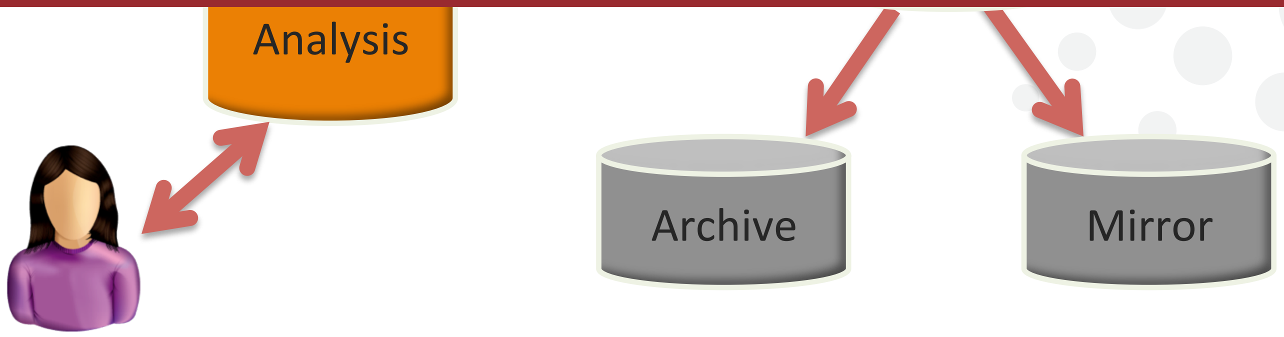
In millions of labs worldwide, researchers struggle with massive data, advanced software, complex

rting

Data movement is a frequent challenge

- Between facilities, archives, researchers
- Many files, large data volumes
- With security, reliability, performance

Next-gen genome sequencer



Accelerate discovery and innovation by outsourcing difficult tasks

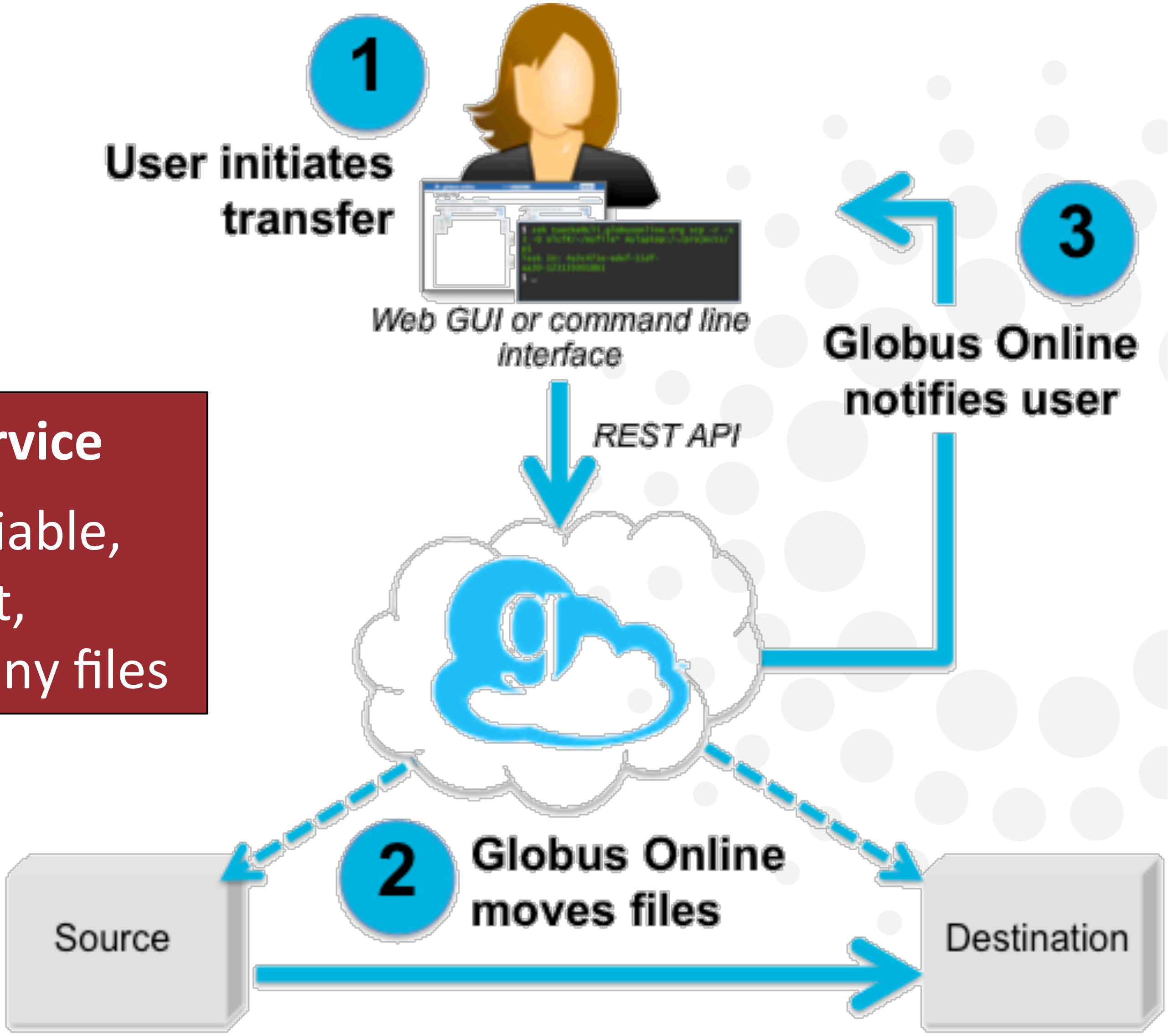
A first step: Automated file movement

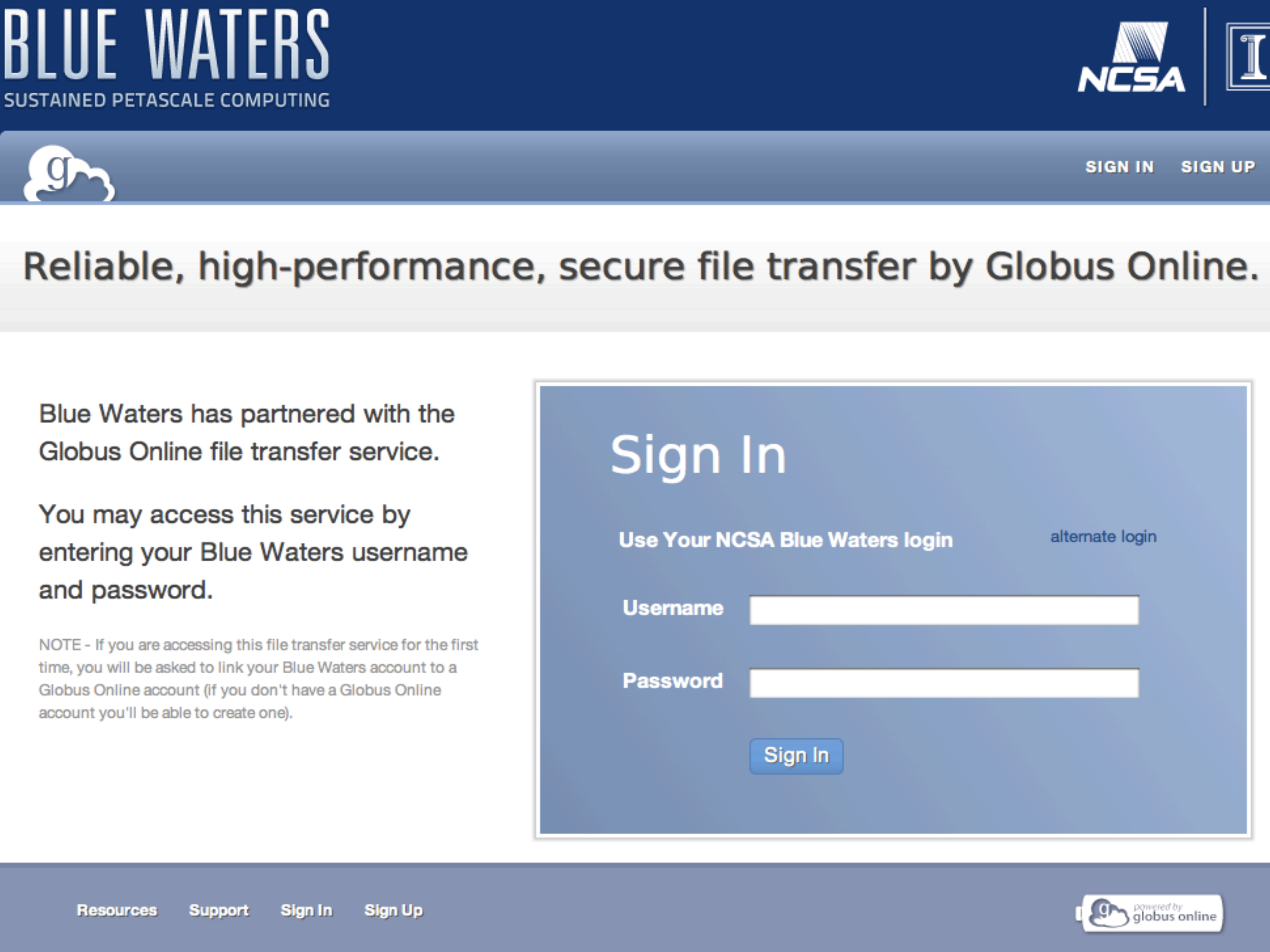
6,000 users
500 M files moved
8 PB moved
100s of endpoints
99.9% availability

File movement as a service

Secure, automated, reliable, high-speed movement, synchronization of many files

Biggest: 1 PB
Most files: 10M
Longest: 100 days
Furthest: Australia


1
User initiates transfer
Web GUI or command line interface
REST API
3
Globus Online notifies user
2
Globus Online moves files
Source
Destination



BLUE WATERS

SUSTAINED PETASCALE COMPUTING

NCSA

SIGN IN SIGN UP

Reliable, high-performance, secure file transfer by Globus Online.

Blue Waters has partnered with the Globus Online file transfer service.

You may access this service by entering your Blue Waters username and password.

NOTE - If you are accessing this file transfer service for the first time, you will be asked to link your Blue Waters account to a Globus Online account (if you don't have a Globus Online account you'll be able to create one).

Sign In

Use Your NCSA Blue Waters login alternate login

Username

Password

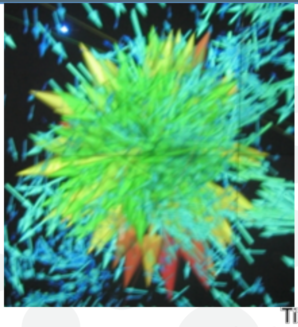
Sign In

Resources Support Sign In Sign Up

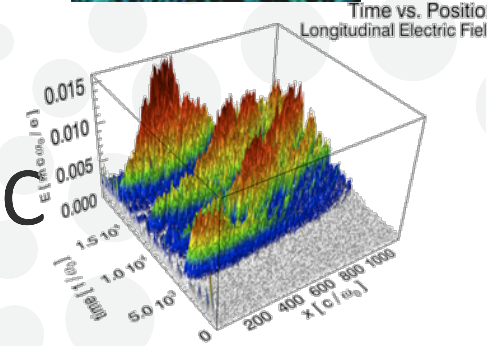
powered by globus online

Examples of Globus Online in action

- K. Heitmann (ANL) moves 22TB **cosmology** data at 5 Gb/s LANL → ANL



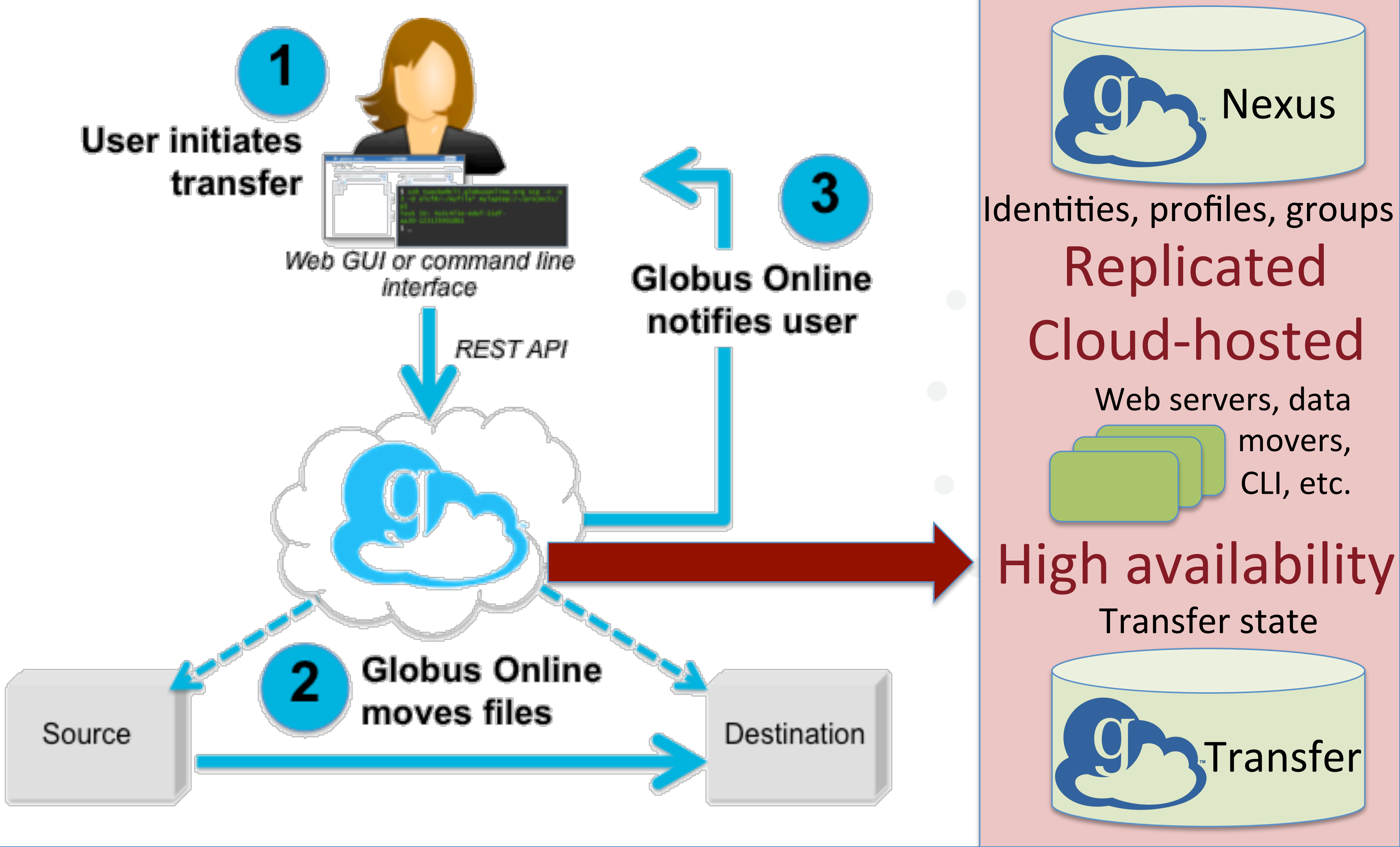
- B. Winjum (UCLA) moves 900K-file **plasma physics** datasets UCLA - NERSC



- Dan Kozak (Caltech) replicates 1 PB **LIGO astronomy** data for resilience

- Recommended by many supercomputer centers, genome facilities, light sources, universities

Globus Online under the covers



1 User initiates transfer

Web GUI or command line interface

REST API

3 Globus Online notifies user

2 Globus Online moves files

Source

Destination

Nexus

Identities, profiles, groups

Replicated

Cloud-hosted

Web servers, data movers, CLI, etc.

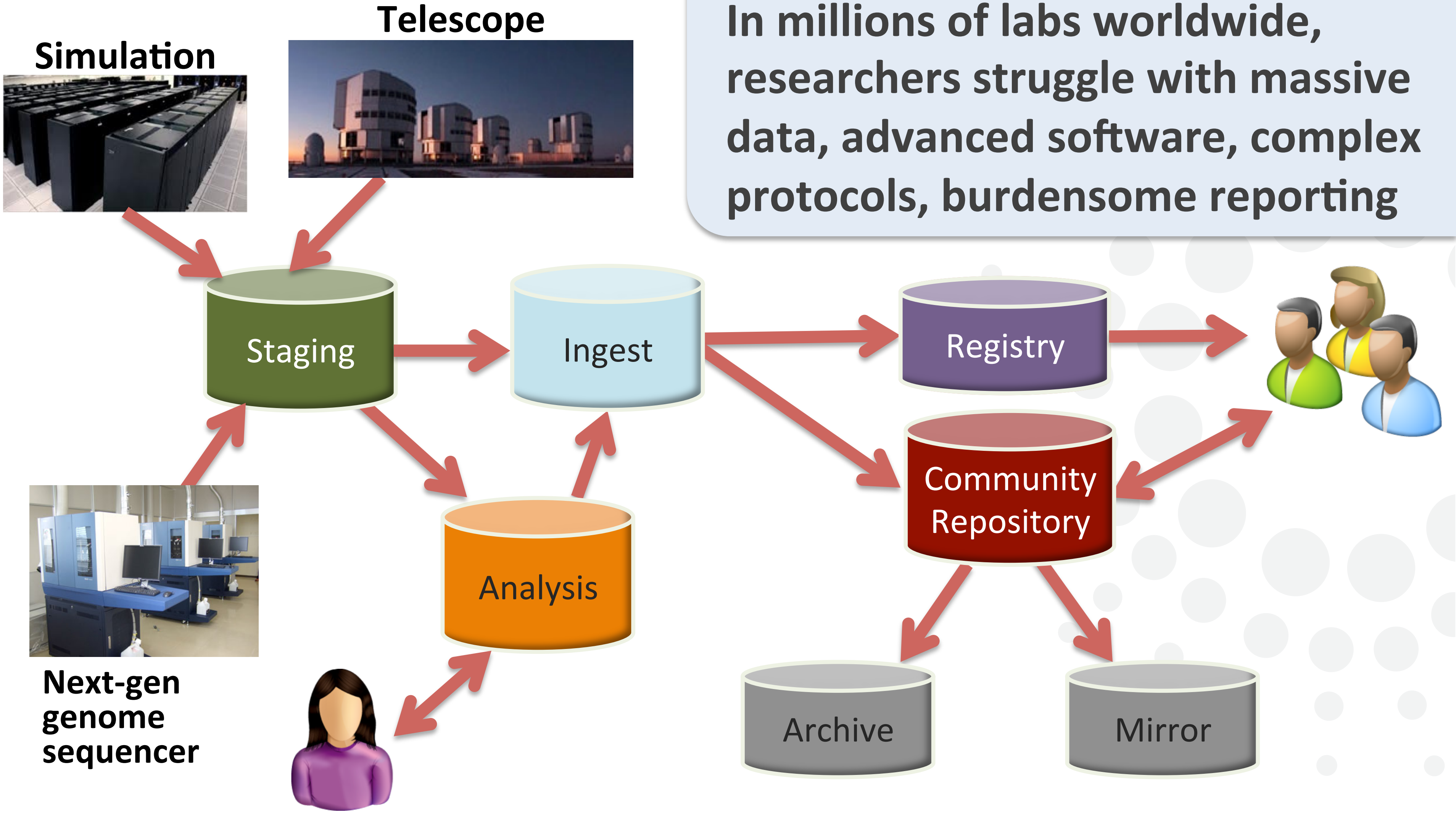
High availability

Transfer state

Transfer

Need much more than file movement

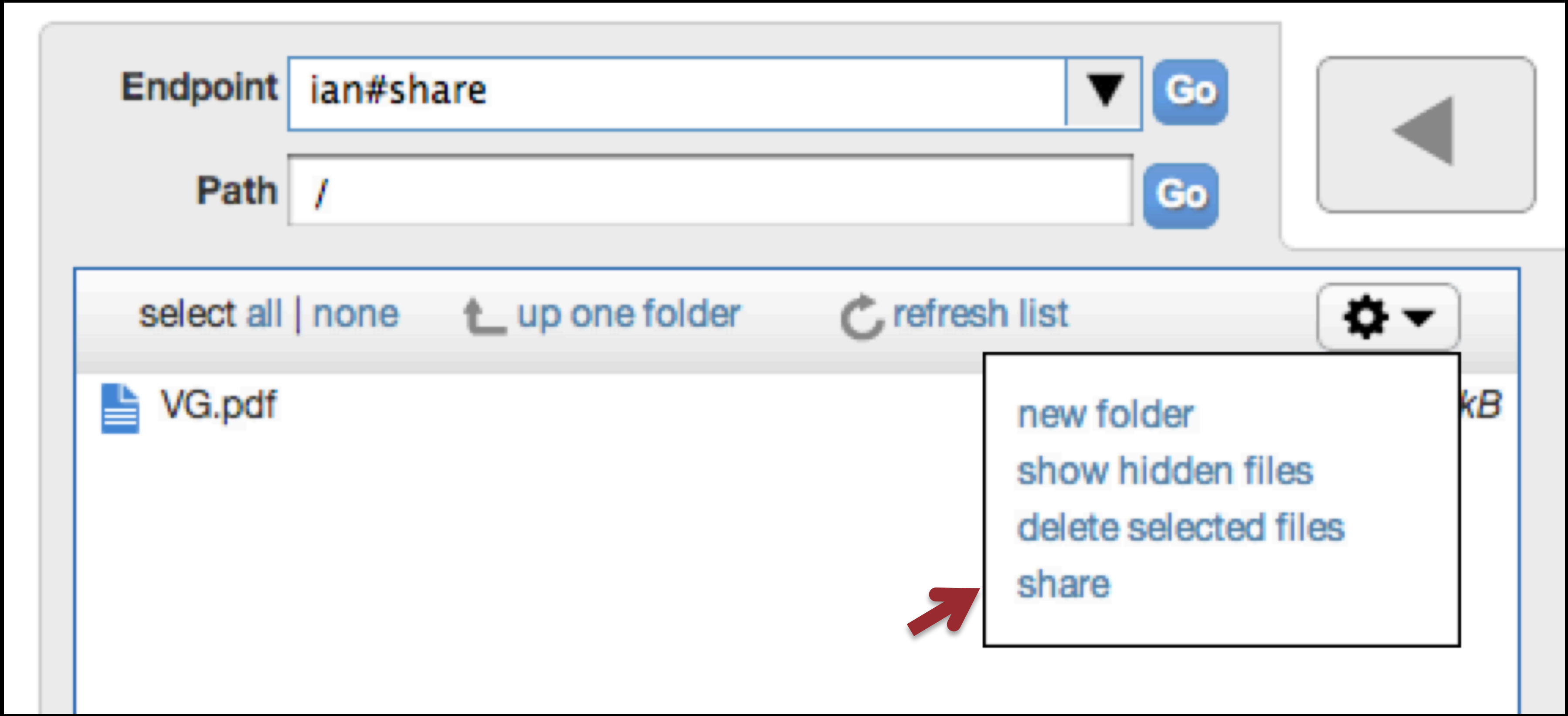
In millions of labs worldwide, researchers struggle with massive data, advanced software, complex protocols, burdensome reporting



Accelerate discovery and innovation by outsourcing difficult tasks

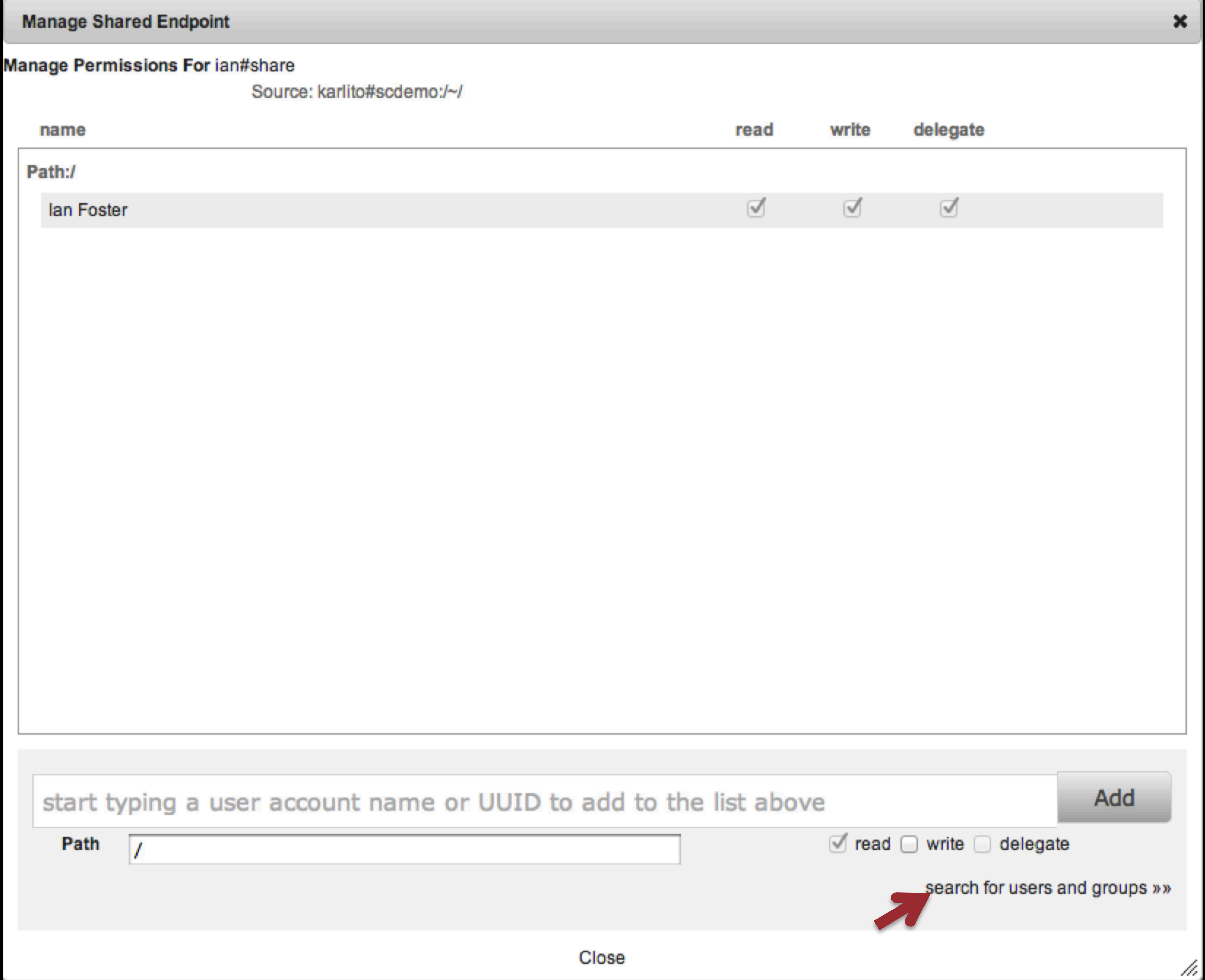
We are also working on other activities

- Sharing and distribution
  - User-managed sharing with individuals and groups
- Ingest and publication
  - Imagine a DropBox that not only replicates, but also extracts metadata, catalogs, converts
- Cataloging
  - Virtual views of data based on user-defined and/or automatically extracted metadata
- Computation
  - Associate computational procedures, orchestrate application, catalog results, record provenance



Endpoint ian#share

Path /



Manage Shared Endpoint

Manage Permissions For ian#share

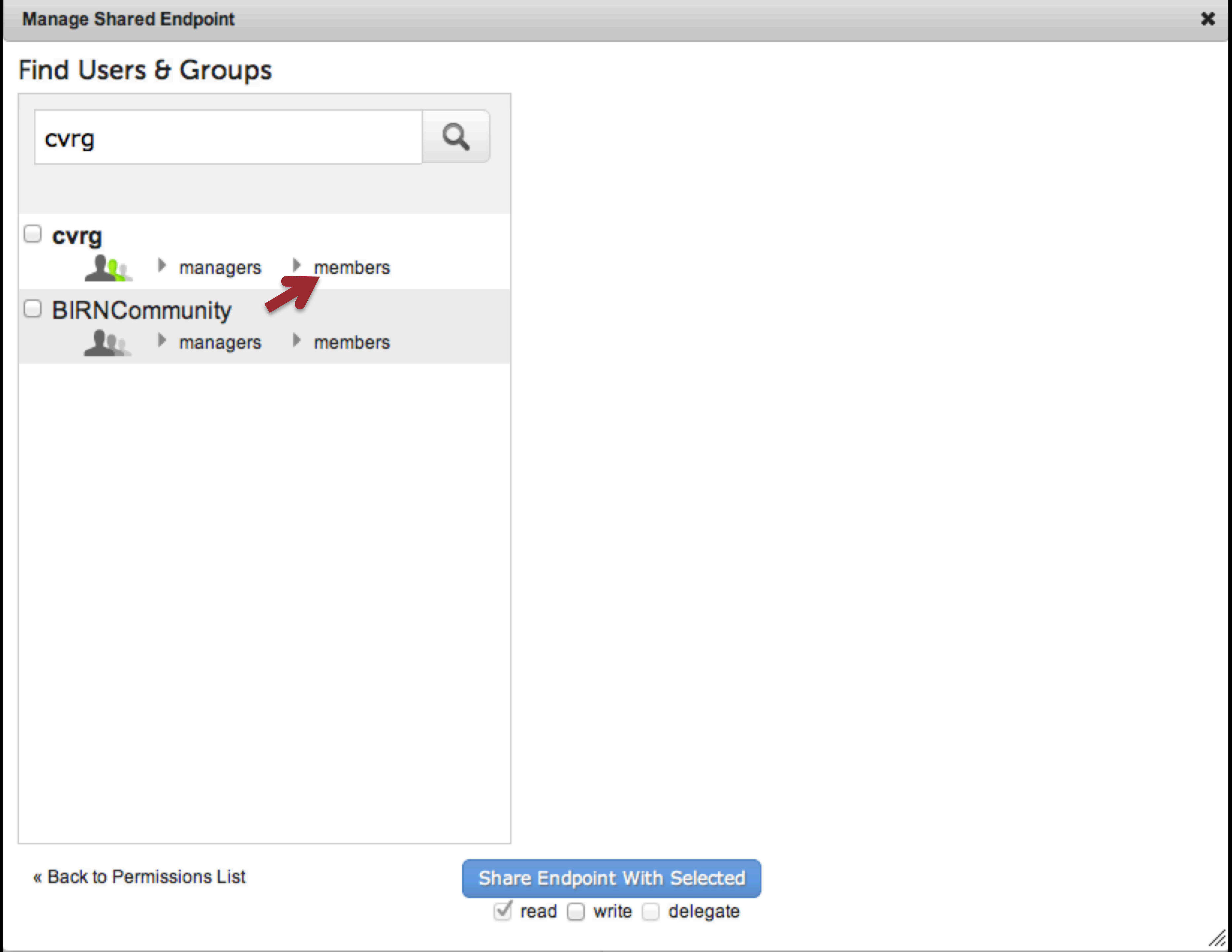
Source: karlito#scdemo:/~/

name	read	write	delegate
Path:/			
Ian Foster	☑	☑	☑

start typing a user account name or UUID to add to the list above

Path /

☑ read □ write □ delegate



Manage Shared Endpoint ✖

Find Users & Groups

cvrg 🔍

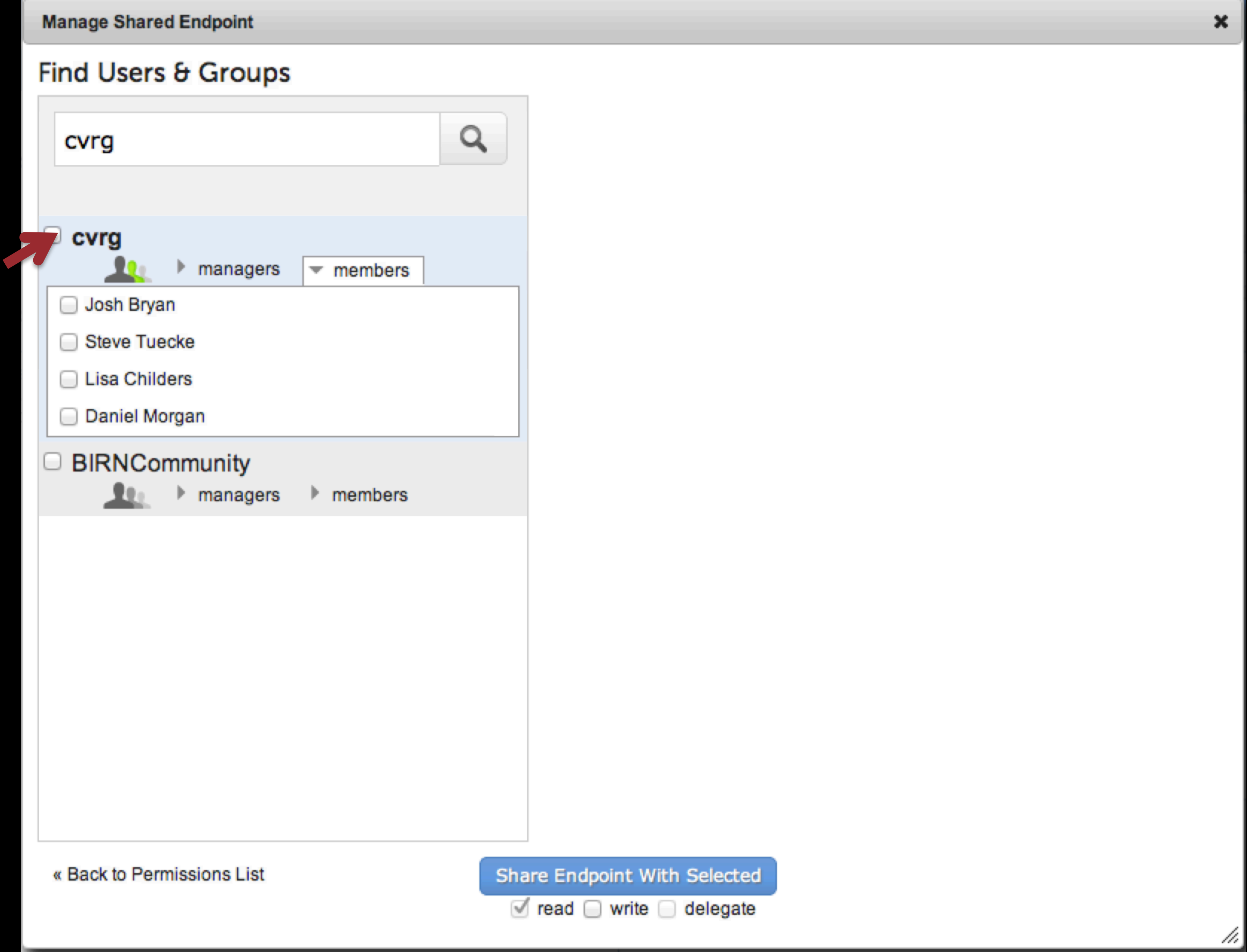
□ cvrg
▸ managers ▸ members

□ BIRNCommunity
▸ managers ▸ members

« Back to Permissions List

Share Endpoint With Selected

☑ read □ write □ delegate



Manage Shared Endpoint ✖

Find Users & Groups

cvrg 🔍

□ **cvrg**

▸ managers ▾ members

□ Josh Bryan

□ Steve Tuecke

□ Lisa Childers

□ Daniel Morgan

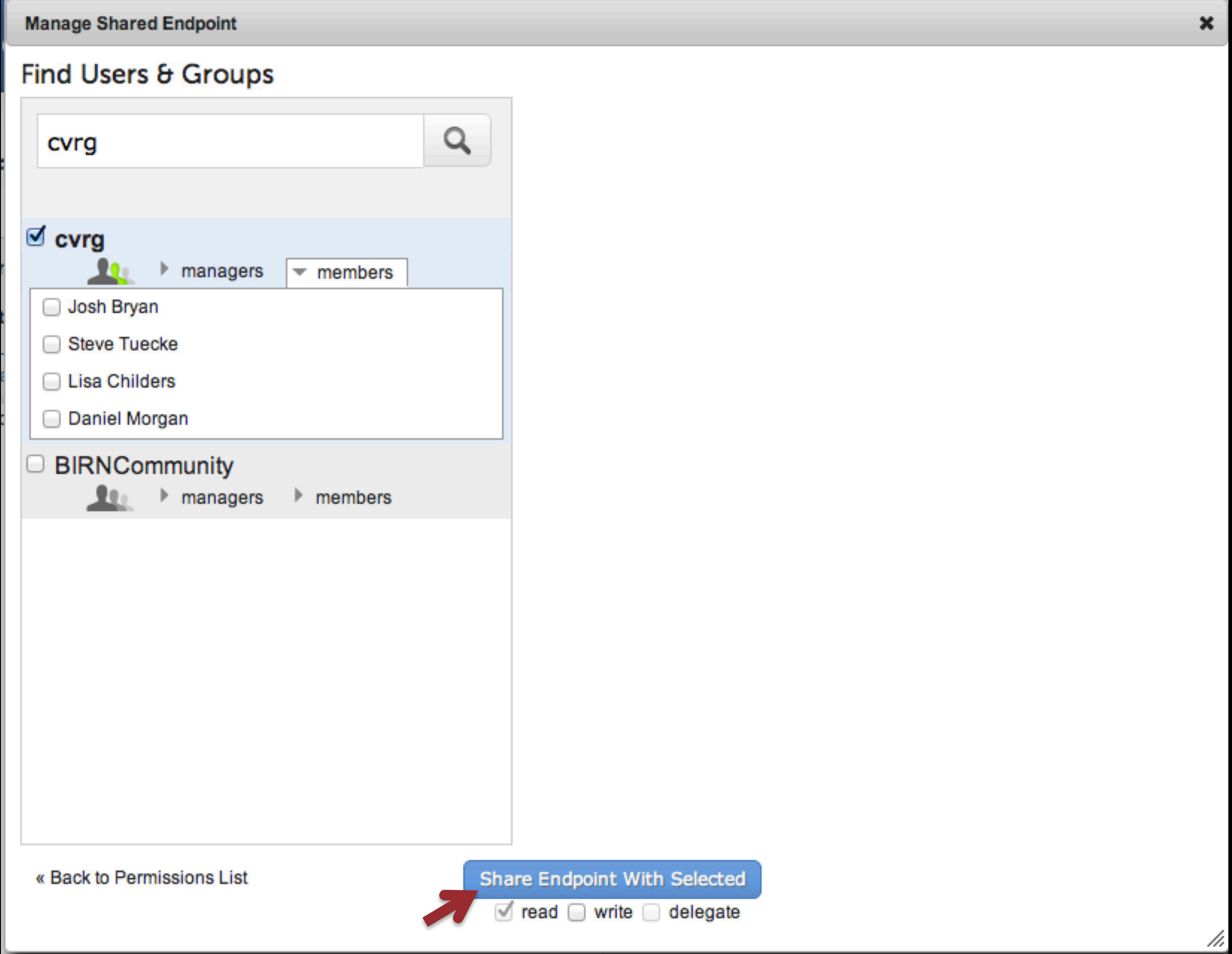
□ BIRNCommunity

▸ managers ▸ members

« Back to Permissions List

Share Endpoint With Selected

☑ read □ write □ delegate



Manage Shared Endpoint ✖

Find Users & Groups

cvrg 🔍

☑ cvrg

▸ managers ▾ members

- ☐ Josh Bryan
- ☐ Steve Tuecke
- ☐ Lisa Childers
- ☐ Daniel Morgan

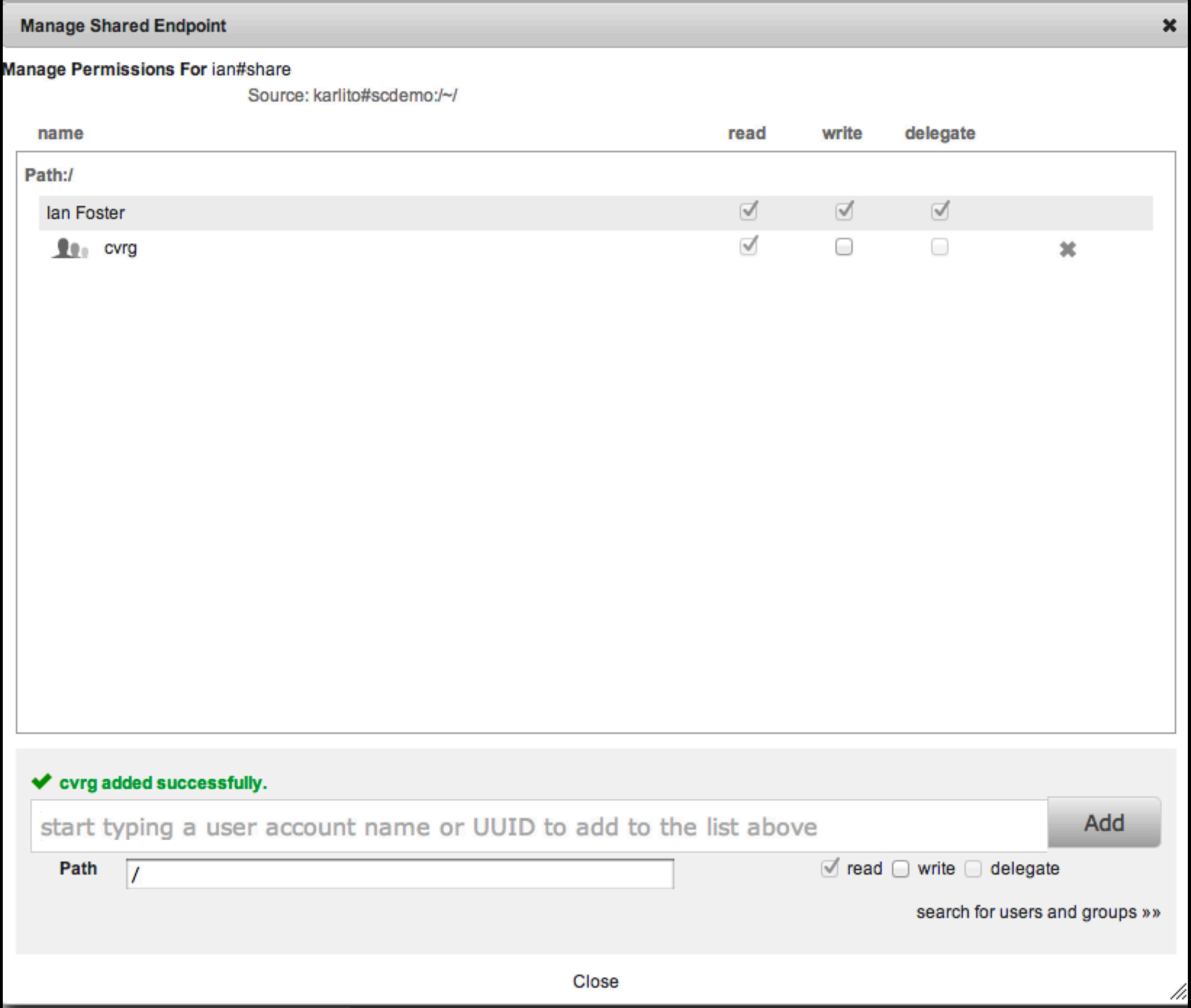
☐ BIRNCommunity

▸ managers ▸ members

« Back to Permissions List

Share Endpoint With Selected

☑ read ☐ write ☐ delegate



Manage Shared Endpoint

Manage Permissions For ian#share

Source: karlito#scdemo:/~/

name	read	write	delegate	
Path:/				
Ian Foster	☑	☑	☑	
cvrg	☑	□	□	✖

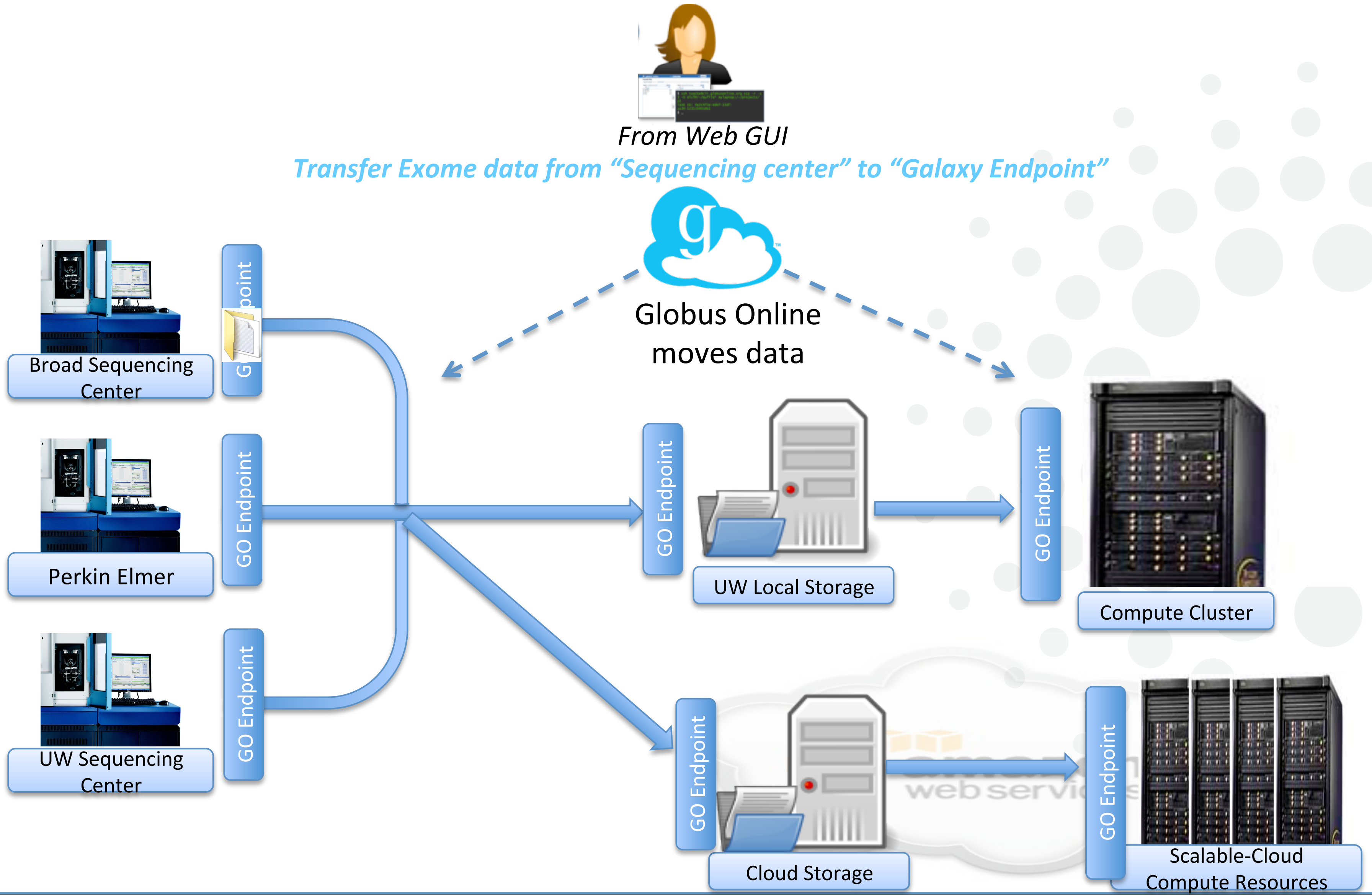
✓ cvrg added successfully.

start typing a user account name or UUID to add to the list above

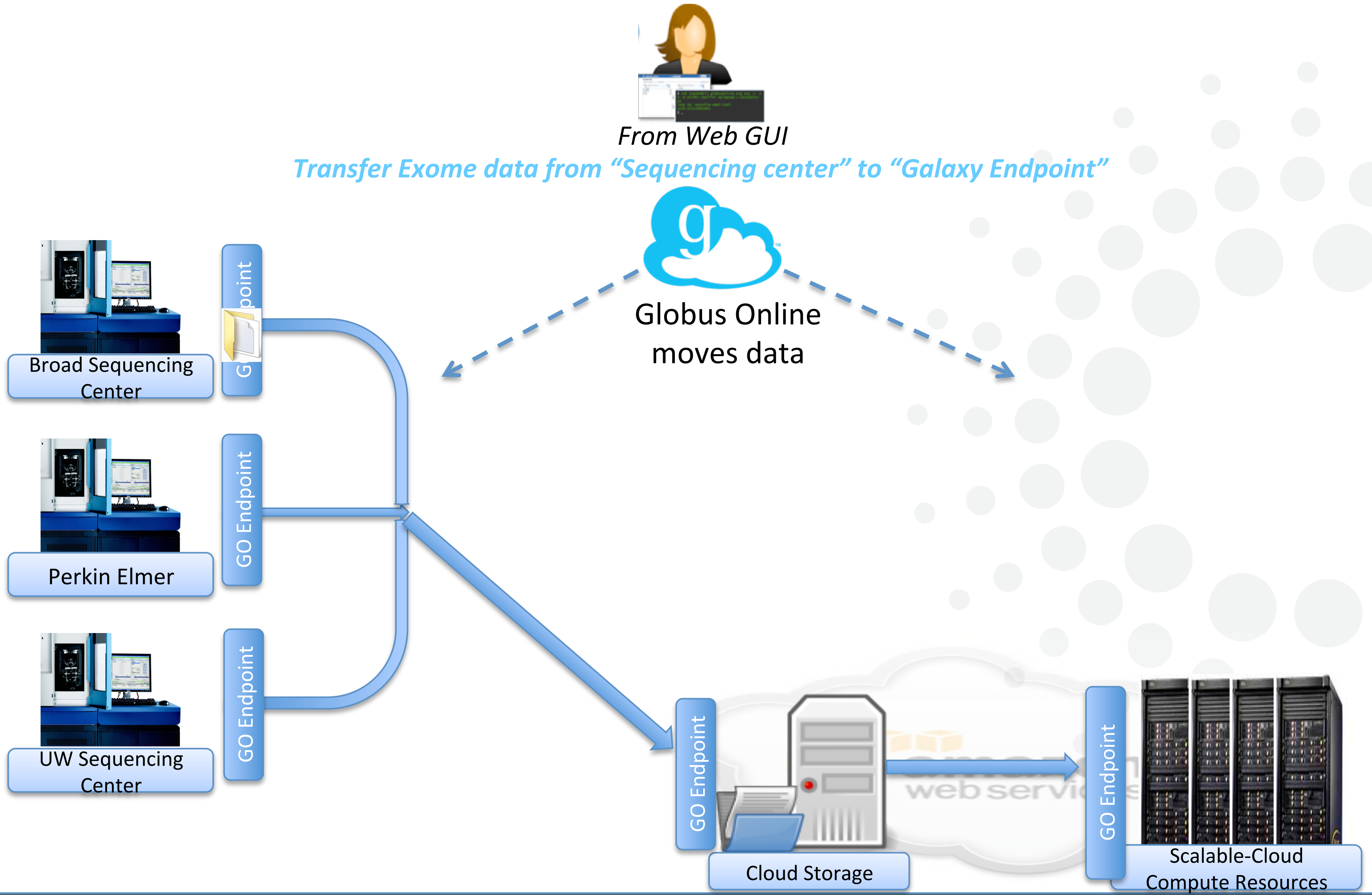
Path: /

☑ read □ write □ delegate

Cloud-based data analysis



Cloud-based data analysis



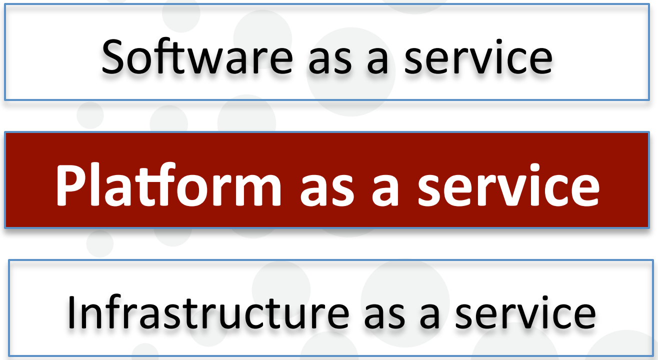
Outsourcing data analysis

- We need **active data repositories** providing for:
  - Integration and maintenance of large, diverse data
  - Not just access to individual elements but demanding user-defined computations over entire datasets
  - Linking of diverse computational results produced by different methods and people
  - Maintenance in the face of diverse updates
- Observations and questions
  - Architecture TBD. Surely HDFS is inadequate?
  - ADRs will surely be heterogeneous
  - Integration with investigator research platforms

A research data management platform

Platform: “technology that enables the creation of products and processes”

- File, catalog, analysis management
- **Identity hub**: Manage user profiles, identities, delegated credentials
- **Group hub**: Manage group membership information



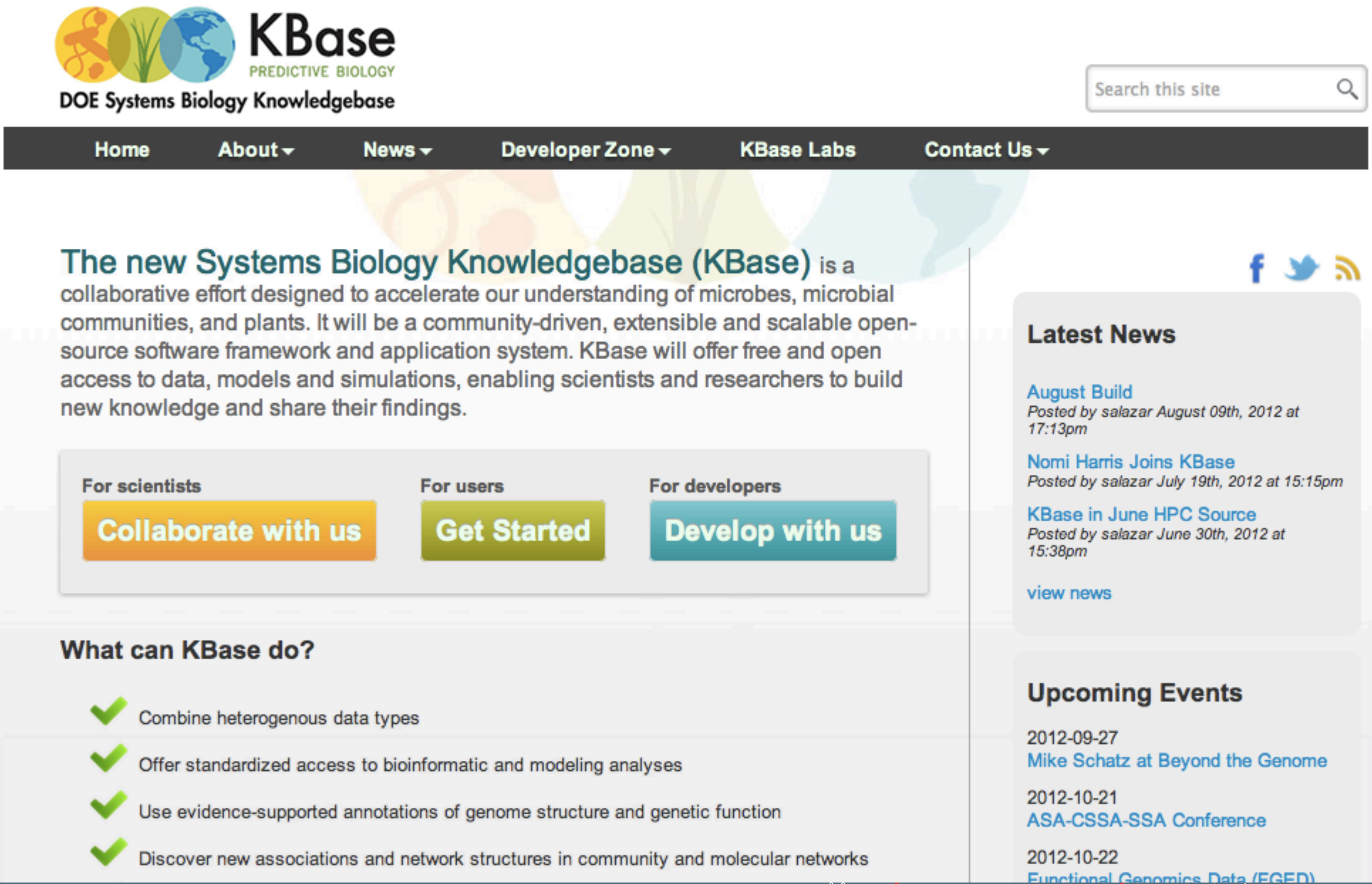
Software as a service

Platform as a service

Infrastructure as a service

All accessible via REST APIs (and CLI), allowing user or team services to offload these functions

Kbase: Identity, group, file movement



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

The new Systems Biology Knowledgebase (KBase) is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It will be a community-driven, extensible and scalable open-source software framework and application system. KBase will offer free and open access to data, models and simulations, enabling scientists and researchers to build new knowledge and share their findings.

What can KBase do?

- ✓ Combine heterogenous data types
- ✓ Offer standardized access to bioinformatic and modeling analyses
- ✓ Use evidence-supported annotations of genome structure and genetic function
- ✓ Discover new associations and network structures in community and molecular networks

Latest News

August Build
Posted by salazar August 09th, 2012 at 17:13pm

Nomi Harris Joins KBase
Posted by salazar July 19th, 2012 at 15:15pm

KBase in June HPC Source
Posted by salazar June 30th, 2012 at 15:38pm

view news

Upcoming Events

2012-09-27
Mike Schatz at Beyond the Genome

2012-10-21
ASA-CSSA-SSA Conference

2012-10-22
Functional Genomics Data (FGED)

kbase.science.energy.gov

In summary: **Big process for big data**

Accelerate discovery and innovation worldwide by providing **research IT as a service**

Outsource time-consuming tasks to

- provide large numbers of researchers with unprecedented access to powerful tools;
- enable a massive shortening of cycle times in time-consuming research processes; and
- reduce research IT costs via economies of scale

Accelerate existing science; enable new science

The team includes [partial list]

At Argonne and UChicago

- Bryce Allen
- Rachana Ananthakrishnan
- Josh Bryan
- Kyle Chard
- Lisa Childers
- Vytas Cuplinskas
- Paul Davé
- Raj Kettimuthu
- Jack Kordas
- Lukasz Lacinski
- Mattias Lidman
- Ravi Madduri
- Stuart Martin
- Dan Morgan
- JP Navarro
- Gigi Rohder
- Steve Tuecke
- Vas Vasiliadis

Collaborators

- Carl Kesselman
- Karl Czajkowski
- Dean Williams
- Francesco de Carlo
- Jim Basney
- Martin Swany
- David Skinner

Opportunities for collaboration

"Research IT as a service: designing and creating the research cloud." For example:

- End-to-end understanding and optimization of local and wide area file transfers
- Identifying research data management activities suitable for outsourcing
- Data-intensive applications involving simulation and/or experimental data
- Architecture and implementation of active data repositories

Computation Institute

Thank you!

foster@anl.gov, @ianfoster

www.ci.anl.gov

www.mcs.anl.gov

www.globusonline.org, @globusonline