



THE UNIVERSITY OF  
**CHICAGO**

**Research  
Computing  
Center**

# UChicago Campus support for research data management

H. Birali Runesha

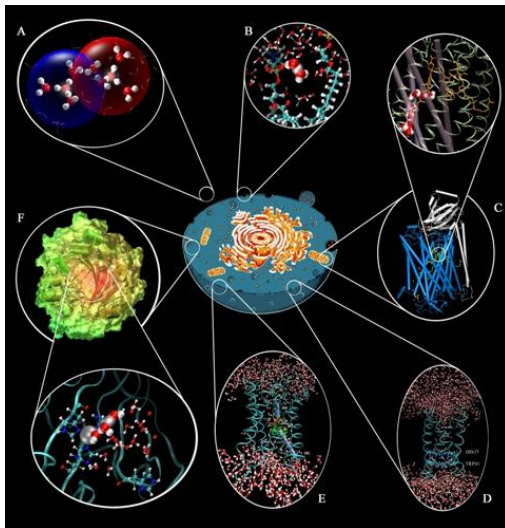
[runesha@uchicago.edu](mailto:runesha@uchicago.edu)

GlobusWORLD Conference, April 18, 2013

---

# Research at UChicago

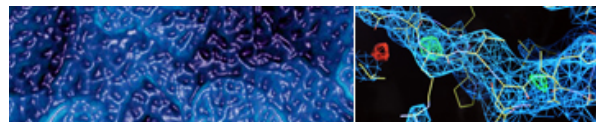
- 4 Divisions and the College: Physical Sciences, Biological Sciences, Social Sciences and Humanities.
- 19 Institutes and dozens centers



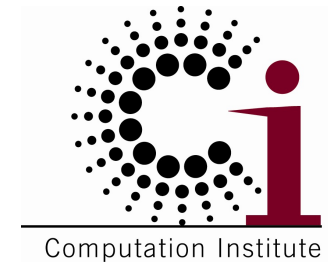
Institute for  
Biophysical Dynamics



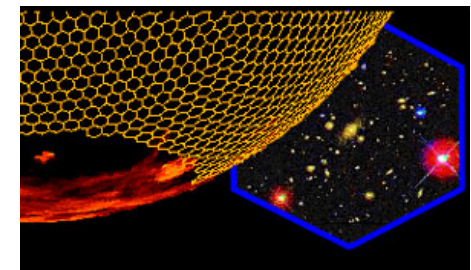
Enrico Fermi Institute



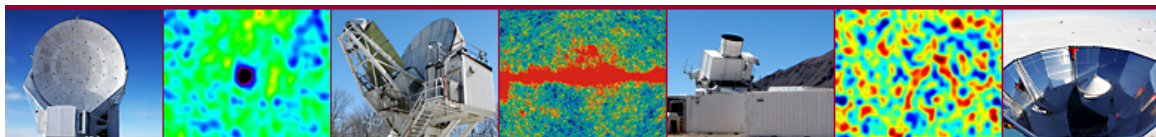
Institute for Molecular Engineering



Computation Institute



Joint Institute for Nuclear Astrophysics



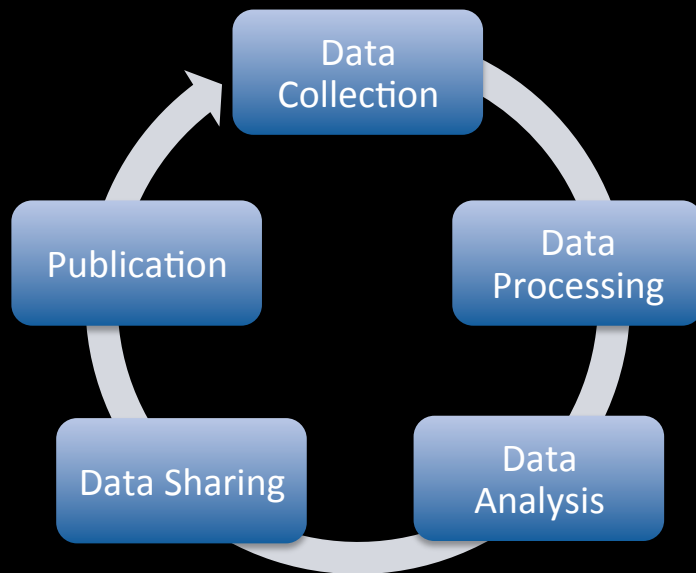
Kavli Institute for Cosmological Physics



Institute for  
Genomics &  
Systems Biology

A wide array of research disciplines  
with various computation and data  
management needs

# Research Data: A challenge for researchers



- Data from experiments, observations and simulations
- Data- and compute-intensive, Integrative, multiscale
- Multi-disciplinary Collaborations
- Individuals, groups, teams, communities

Poor data management  
Power-inefficient hardware  
Inadequate software, ad hoc solutions  
Lack of skilled IT staff  
Poor performance, excessive costs





# Some of the challenges

- Research instrumentation creates a flood of data
  - fMRI, DTI, NMR, gene sequencing, x-ray crystallography
  - computer simulations create an even larger tide of data
- Data storage and compute resources are highly distributed
- Storing of data on inadequate media with no consistent back up mechanism.
- Volume of data with no where to store the data
- Complexity associated with using the data.
- Metadata and data preservation
- HIPAA, Sensitive data
- Lack of collaborative tools to transport and share data
- Etc.



THE UNIVERSITY OF  
**CHICAGO**

**Research  
Computing  
Center**

---

## Mission

1. Enable research and scholarship by providing access to centrally managed research computing, storage and visualization resources.
2. Provide technical user support, education and training.
3. Deploy advanced technologies to enable research discoveries and innovations.

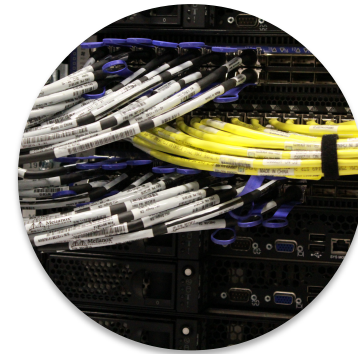
[rcc.uchicago.edu](http://rcc.uchicago.edu)

A vision towards data management and computing as a service.



# RCC Hardware: Compute

- **Tightly Coupled nodes**—FDR10 InfiniBand.
  - Intel Sandy Bridge—16 cores per node
  - 2.6 GHz, 32 GB of memory
  - More than 325 nodes
- **Loosely Coupled nodes**—GigE
- **Shared Memory nodes** —256 GB to 1 TB per node
- **GPU nodes** —NVIDIA M2090 and Kepler
- **Hadoop MapReduce nodes**



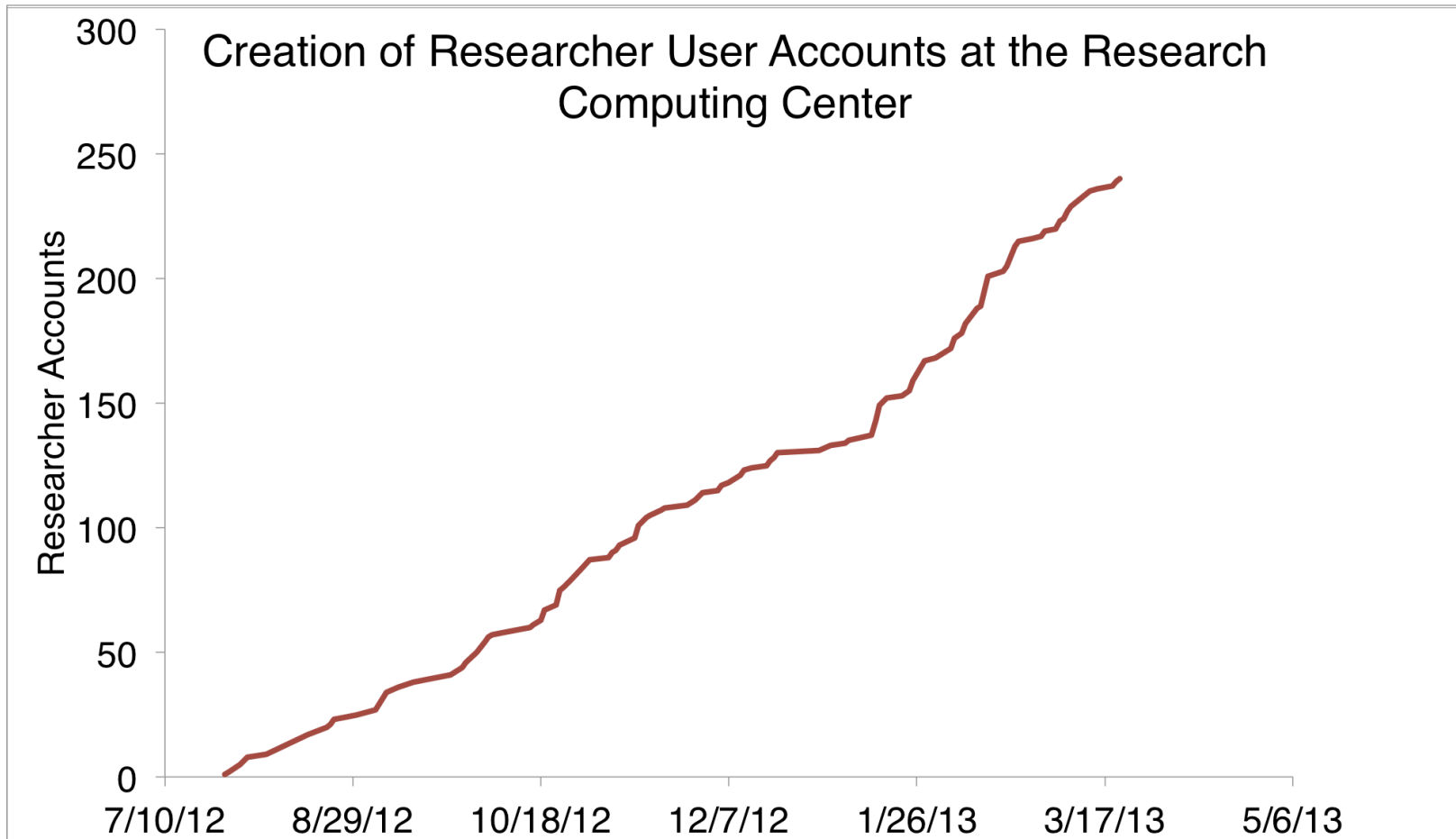
# RCC Hardware: Storage

- **Home**
  - backed up and snapshotted
  - exclusive
  - limited storage
- **Capacity (Project)**—: 0.5PB – 1.5 PB
  - backed up and snapshotted
  - shared with group
  - scalable to many TB
- **High Performance (/scratch)**—75TB
  - time-limited
  - not backed up
- **Tape Backup**
  - Home and project backed up and snapshotted

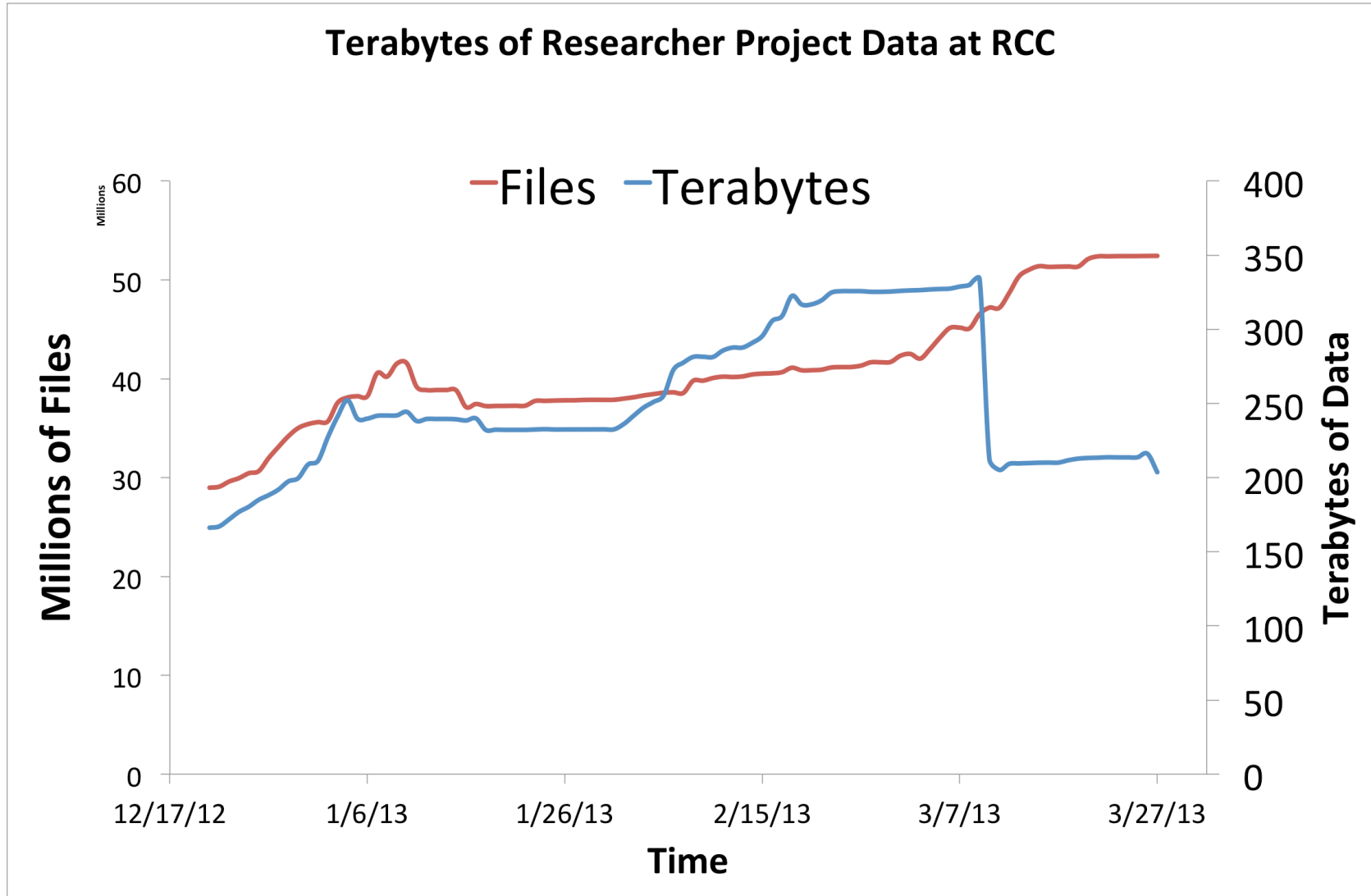


RCC is approaching ~1.5 PB of storage

# Growth Phase



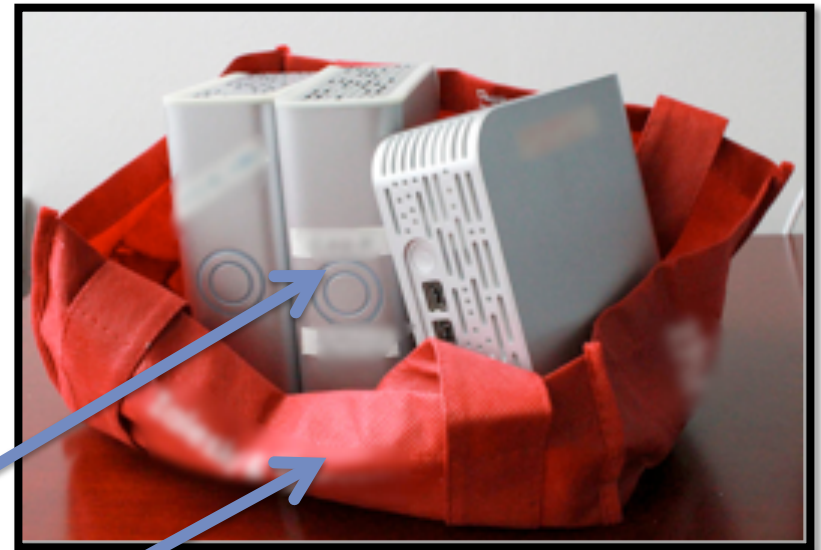
# Growth Phase



# RCC's first data transfer

“Currently we rely far too heavily on personal external hard disks which can be flaky.

Do you have any suggestions?”



Capacity Project Storage

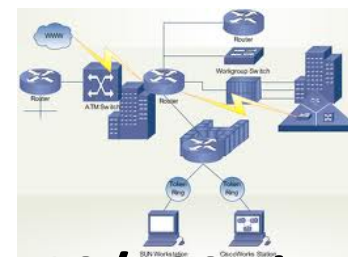
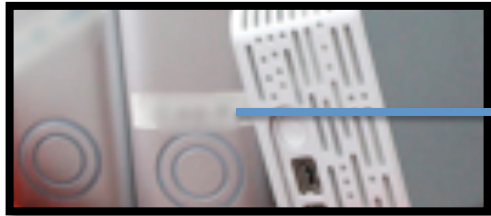
Data Transfer Utility

Network



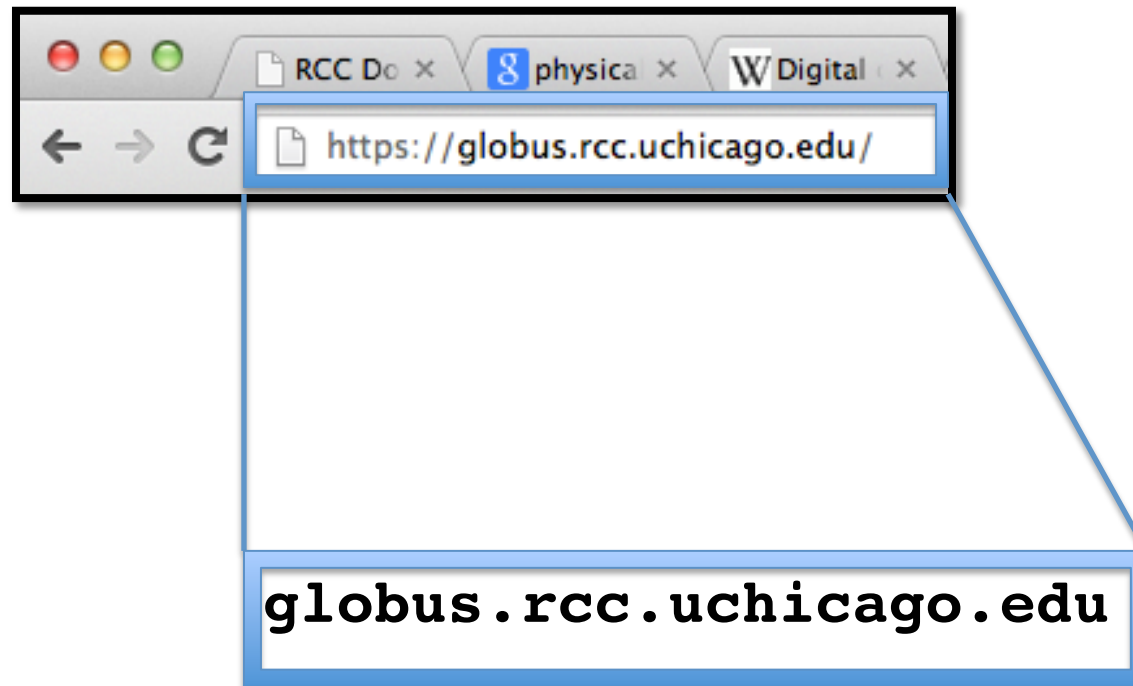


It was clear we had a long way to go...



**10/100 gigabit**

# How is Globus Online implemented at RCC?







## Transfer Activity

Cancel

◀◀ 1 of 1 ▶▶

View 25 Records

	Status	Label	Task Progress	Completion Time	Request Time
<input type="checkbox"/>	✓	Task Id:e6f1fb78-9ca0-11e2-97ce-123139404f2e	1 / 1	04/03/2013 03:56 PM	04/03/2013 03:56 PM
<input type="checkbox"/>	✓	Task Id:f379e4d4-d6a1-11e1-bf56-1231380b8963	1 / 1	07/25/2012 04:44 PM	07/25/2012 04:44 PM
<input type="checkbox"/>	✓	Task Id:c5beaa62-d6a0-11e1-bf56-1231380b8963	1 / 1	07/25/2012 04:36 PM	07/25/2012 04:36 PM

Cancel

◀◀ 1 of 1 ▶▶

View 25 Records



*“I have about 1.5TB  
of data I'd like to  
transfer at this time*

*What's the best  
way to transfer  
data?*

“

Stephanie E. Palmer, Ph.D.  
Assistant Professor  
Organismal Biology and Anatomy  
University of Chicago

“

I just set up my  
Globus account and got  
a transfer started.

...and HOLY  
COW! it's FAST!!!

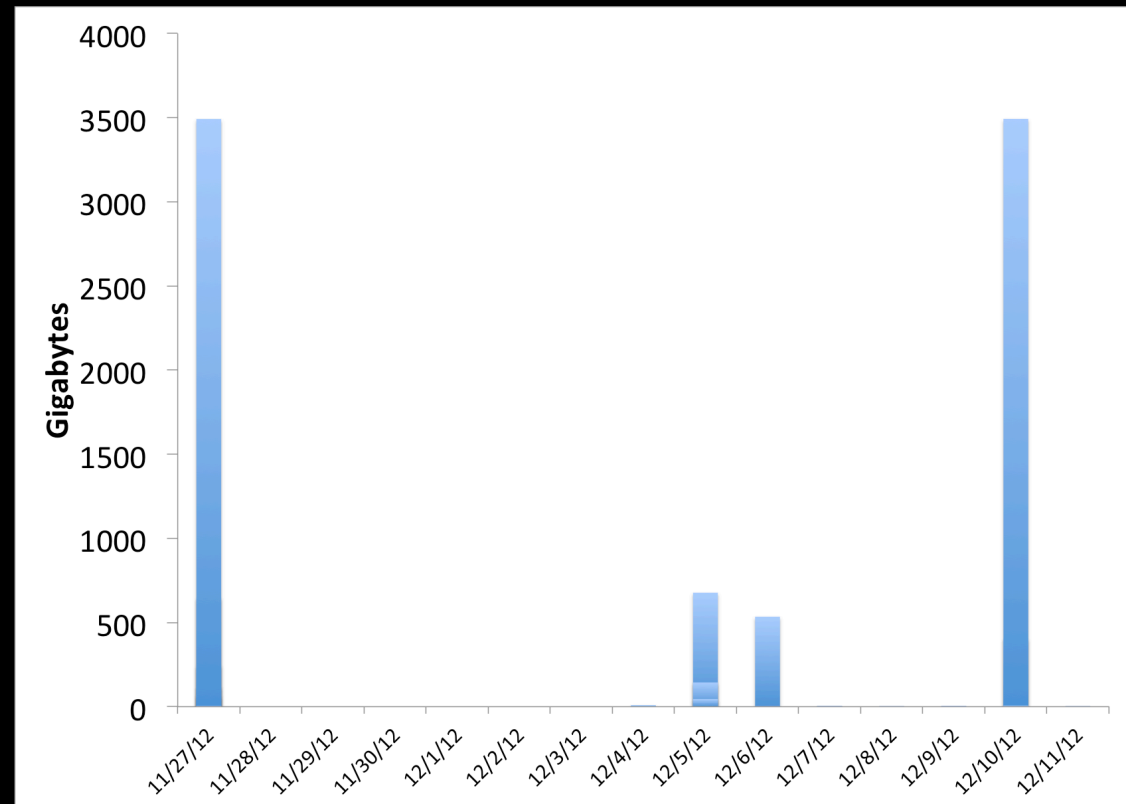
”

Stephanie E. Palmer, Ph.D.  
Assistant Professor  
Organismal Biology and Anatomy  
University of Chicago

# Case Study (Elliott)

*I need to start moving a pretty big chunk of data off of PADS and on to RCC ASAP*

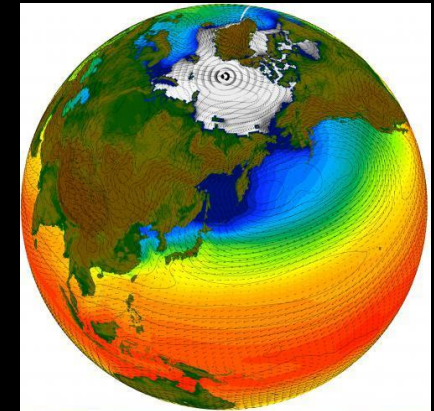
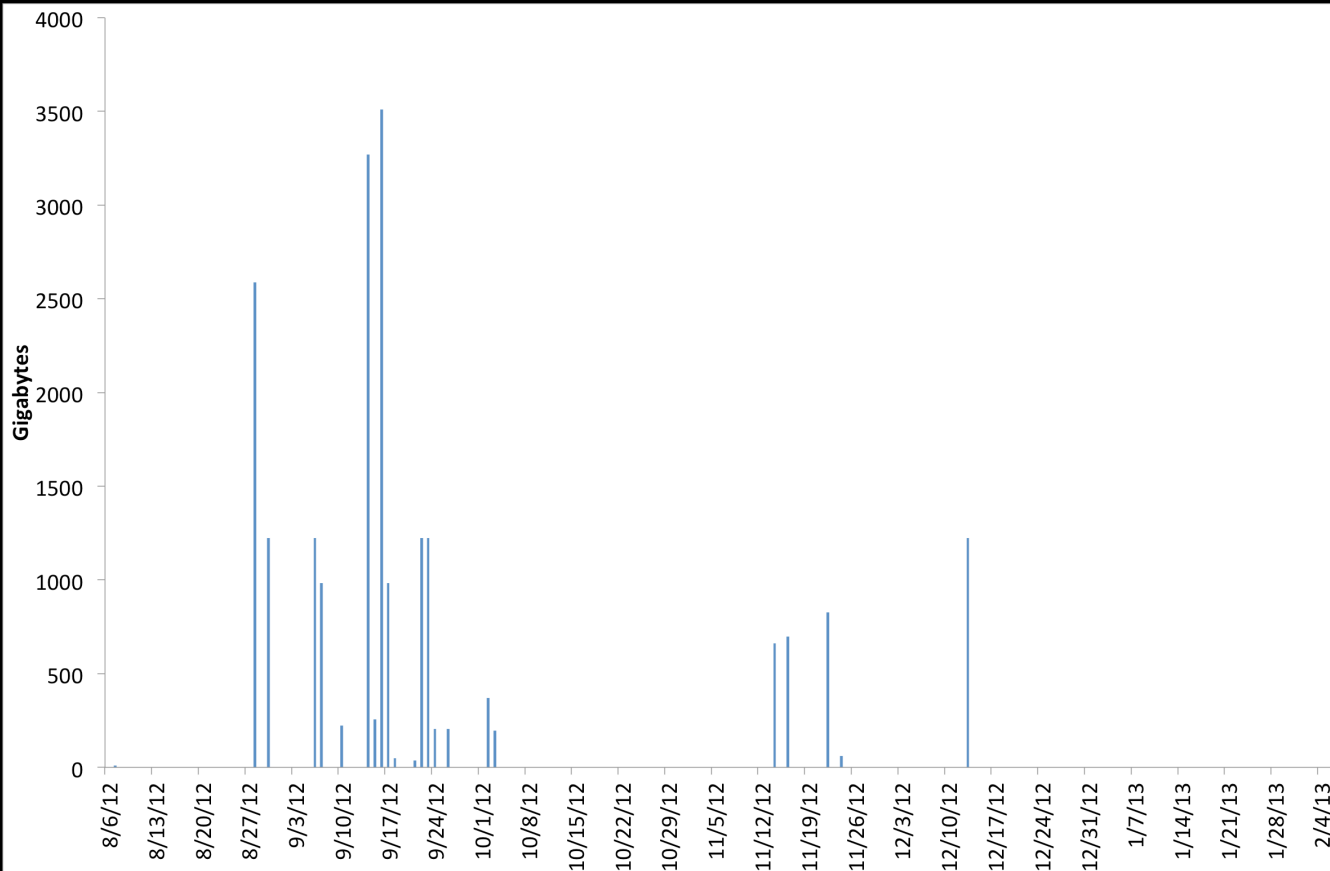
“pretty big chunk” equates to 45 million files and 10 terabytes, in this case





# Case Study (Moyer)


*We have 35 TB on PADS right now and more ... sitting in limbo.*



60 terabytes, 5.5 million files

# Case Study: Matt Becker

Friday, February 3rd, 2012

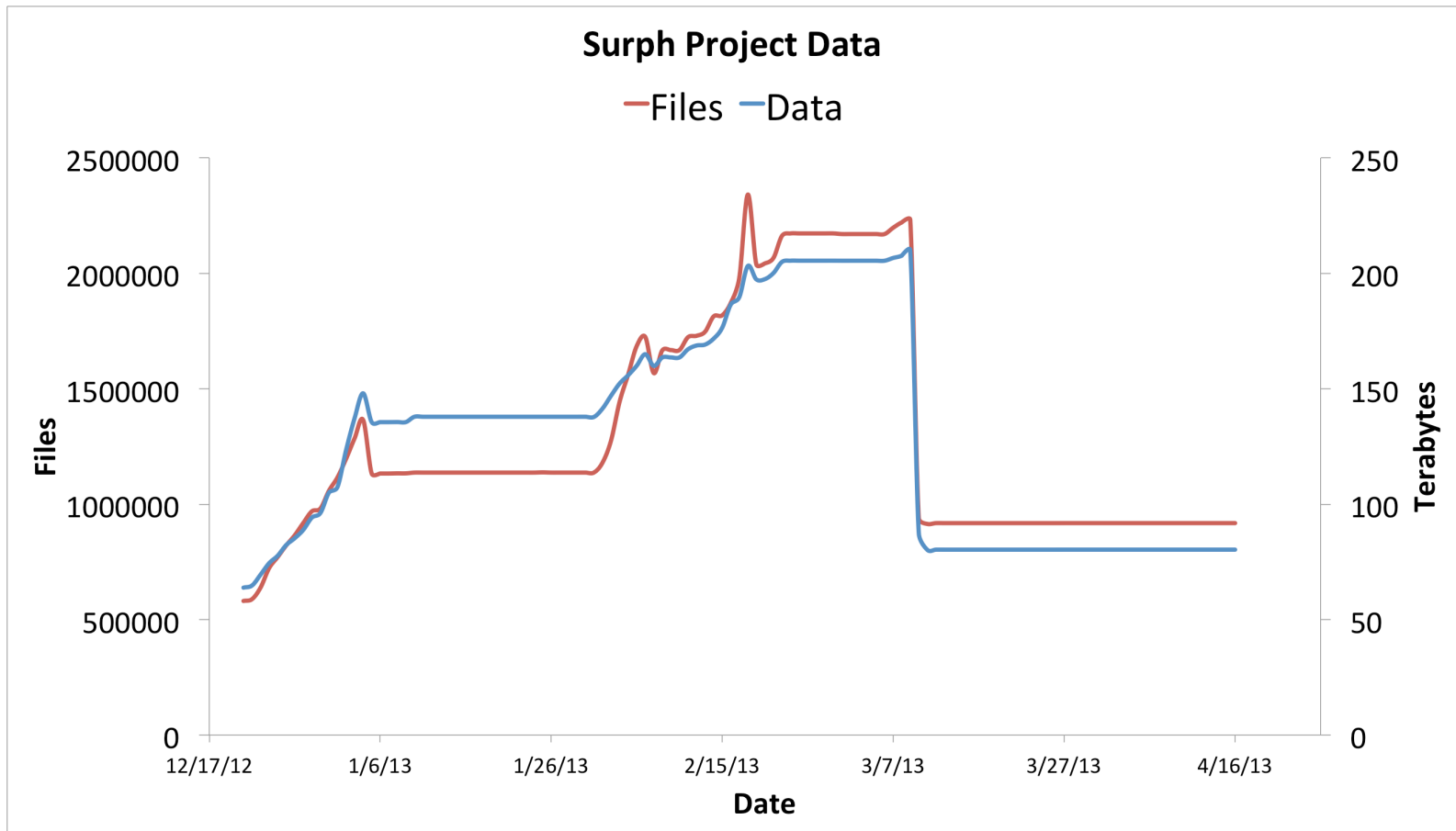
 comments off

## February 2012 User of the Month: Matt Becker

For February's User of the Month, I'm pleased to announce the winner is Matthew Becker from University of Chicago!

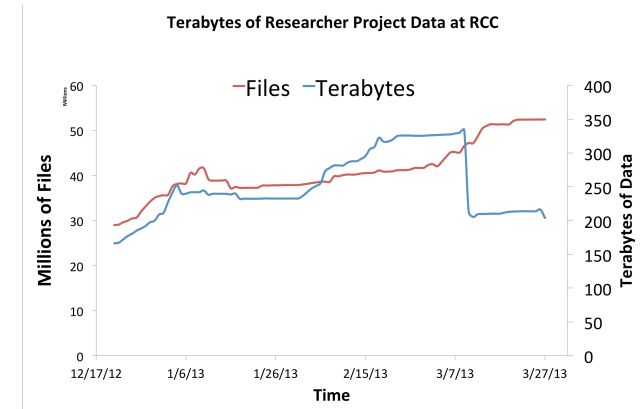
Moving data from/to RCC Midway storage  
to XSEDE resources and Stanford

# Becker's Activity at RCC

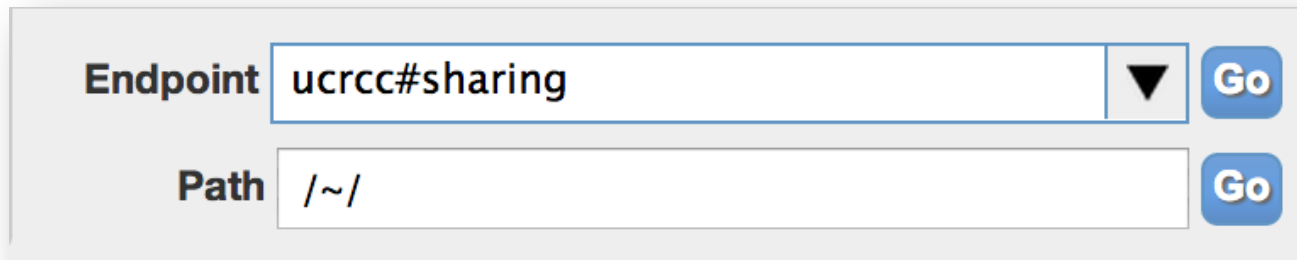


# The Need for Sharing

- RCC is a hub for research projects, providing a 'home' for data that is generated in many locations
- Researchers naturally want to share this data with their offsite collaborators
- Currently sharing is clunky, requiring each user to have a CNetID



# Globus Online Sharing



The image shows a screenshot of the Globus Online Sharing interface. It features two input fields: 'Endpoint' and 'Path'. The 'Endpoint' field contains the text 'ucrcc#sharing' and has a dropdown arrow on its right side. The 'Path' field contains the text '/~/'. Both fields have a blue 'Go' button to their right.

- Globus Sharing empowers researchers to manage data sharing on their own, without intervention of RCC or the need to create CNetIDs for collaborators
- RCC is currently testing GO sharing capabilities.
- All tests have been very successful
- Currently selecting pilot research groups for pilot projects
- Will be accessible through rcc account



Manage Data

Groups

runesha's Profile

Sign Out

## Transfer Files

[View Transfer Activity](#)

[Get Globus Connect](#)

Turn your computer into an endpoint.

Endpoint:   Path:

Endpoint:   Path:

select all | none

	HPL	
	TEST	
	TMP	
	TUTORIALS	
	YDA	
	koronis	Folder
	rcchelp	Folder
	scratch-midway	Folder
	software	Folder
	a.out	692.51 kB
	h2a.15_h2d.30-41k.zip	15.58 MB
	hbr01.pem	1.66 kB
	pbslog	1.8 kB
	regatta.tar.gz	566 MB
	test.f	46 b
	test.f90	46 b

new folder  
show hidden files  
delete selected files  
share

select all | none

	Desktop	Folder
	Documents	Folder
	Downloads	Folder
	Dropbox	Folder
	Final	Folder
	Meeting-Notes	Folder
	Meetings_Notes	Folder
	Movies	Folder
	Music	Folder
	Pictures	Folder
	Public	Folder
	Python	Folder
	RCC_staff_meetings	Folder
	RUNESHA	Folder
	SC12	Folder
	Sites	Folder
	ANL	154 b
	Ben	2.11 kB
	CASC_september2012	10.49 kB
	CHPC	4.6 kB

**Label This Transfer**

This will be displayed in your transfer activity.



My Groups

### Create New Group

Group Name

Group Description

**B** *I* U

- Group Can Be Viewed By
- logged in users (visible to all **University of Chicago - Research Computing Center** members)
  - group members only (private)

Create Group

Cancel

Create New Group »

● admin action required

ources.

le groups that you belong  
admin Queue link to get a

users.

nk at the bottom of the left



# Advantage of Using Globus Online at RCC

- Providing users with high –performance data movement
- No need to register with GO with and RCC account
- End-to-end problem determination
- No need to install and configure GridFTP
- Credential management and security
- No need to babysit and troubleshoot data transfers
- A Web 2.0 user interface for data transfer
- Data movement between facilities, researchers
- Secure and reliable data movement of many files and large data volumes

Thank you

[rcc.uchicago.edu](http://rcc.uchicago.edu)