Adapting federated cyberinfrastructure for shared data collection facilities in structural biology

GlobusWorld 2012

Ian Stokes-Rees ijstokes@seas.harvard.edu @ijstokes
Harvard Medical School

J. Synchrotron Rad. (2012) 19

doi:10.1107/S0909049512009776

computer programs

Journal of
Synchrotron
Radiation

ISSN 0909-0495

Received 9 December 2011
Accepted 5 March 2012

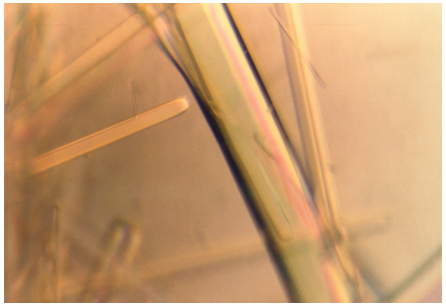
Adapting federated cyberinfrastructure for shared data collection facilities in structural biology

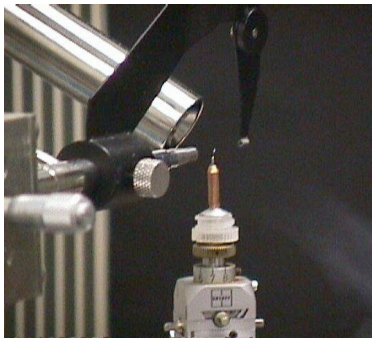
Ian Stokes-Rees,[a] Ian Levesque,[a,b] Frank V. Murphy IV,[c] Wei Yang,[d] Ashley Deacon[d] and Piotr Sliz[a]*

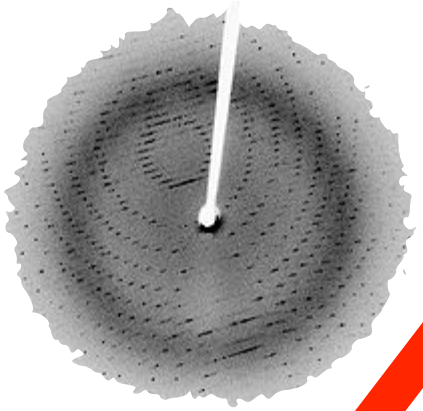
Online 6 April 2012

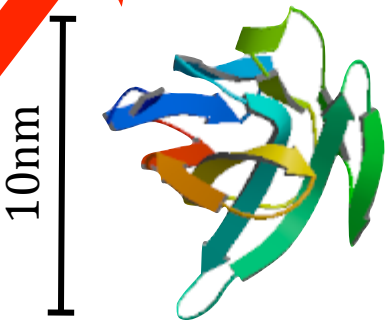
JCSG (U54 GM094586 and GM074898). Portions of this research were performed at the Stanford Synchrotron Radiation Lightsource (SSRL), SLAC National Accelerator Laboratory. The SSRL is a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. The authors would also like to thank Steve Tuecke, Rachana Ananthakrishnan and Raj Kettimuthu from Globus Online for their support and encouragement with this project, and James Withrow from NE-CAT for technical support in the deployment process. Terrence Martin, at the University of California San Diego (UCSD), and Brian Bockleman, at the University of Nebraska-Lincoln, provided invaluable assistance with service configuration and data management on the HadoopFS system at UCSD.

Structural Biology:
Study of Protein Structure and Function





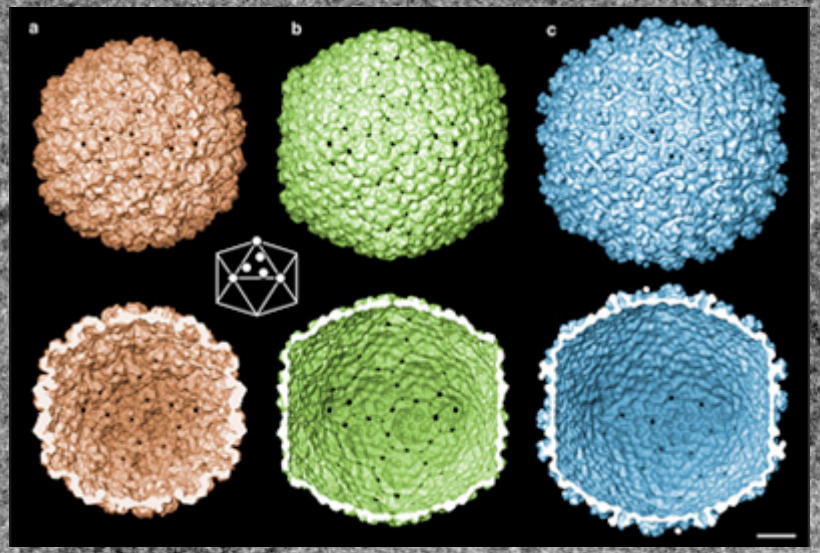


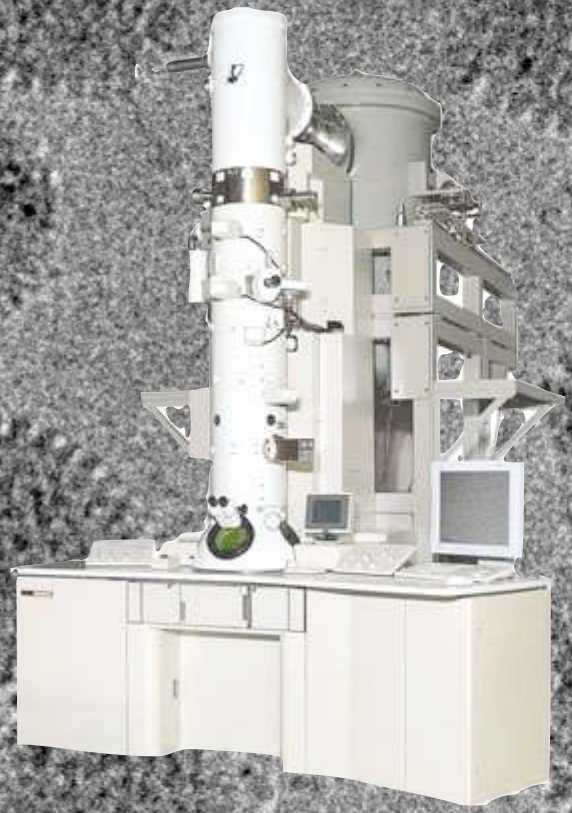


- Shared scientific data collection facility
- Data intensive (10-100 GB/day)

ijstokes@hkl.hms.harvard.edu

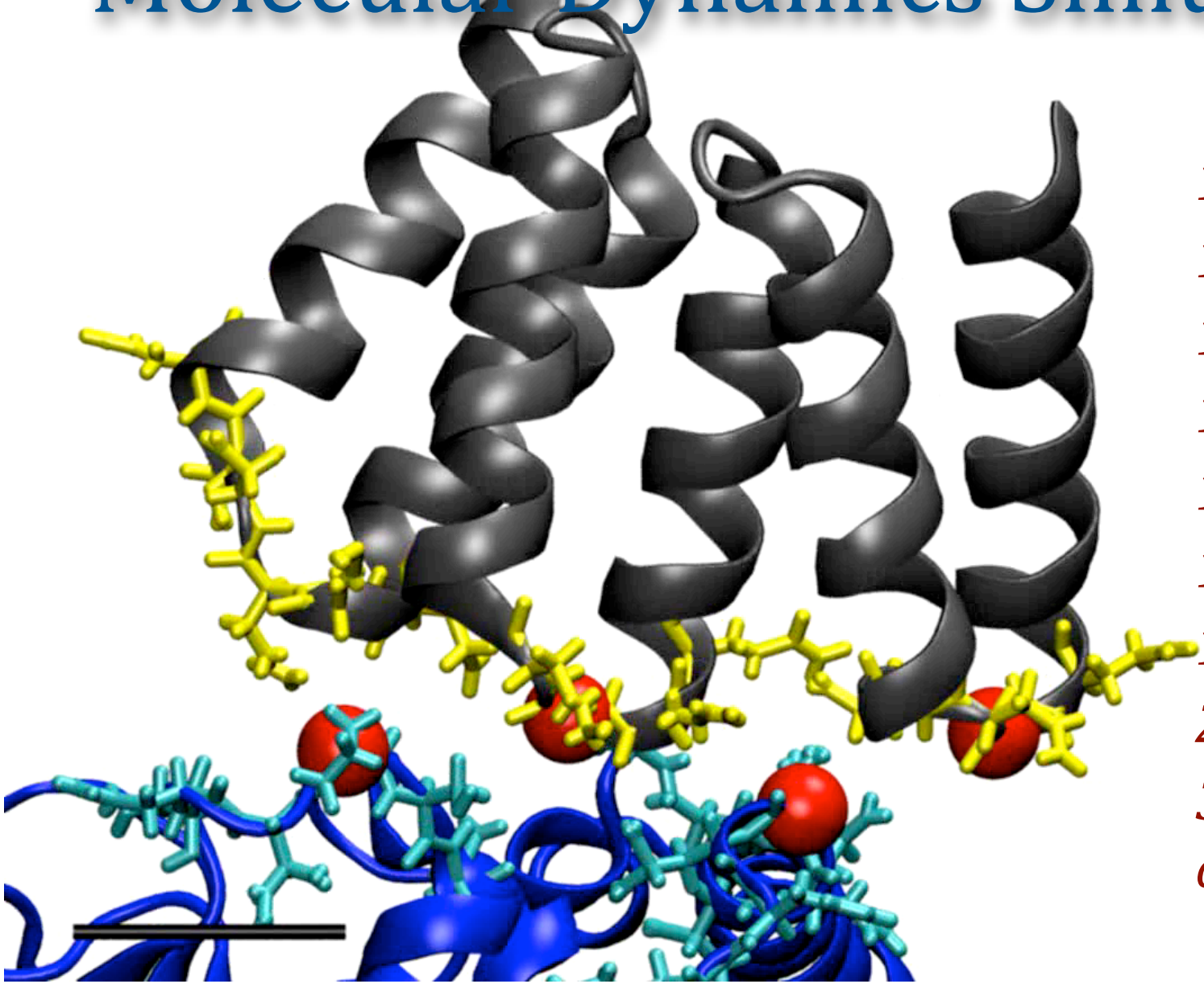
Cryo Electron Microscopy





- *Previously, 1-10,000 images, managed by hand*
- *Now, robotic systems collect millions of hi-res images*
- *estimate 250,000 CPU-hours to reconstruct model*
- *20-50 TB of data in most extreme case*

Molecular Dynamics Simulations



1 fs time step

1ns snapshot

1 us simulation

1e6 steps

1000 frames

10 MB / frame

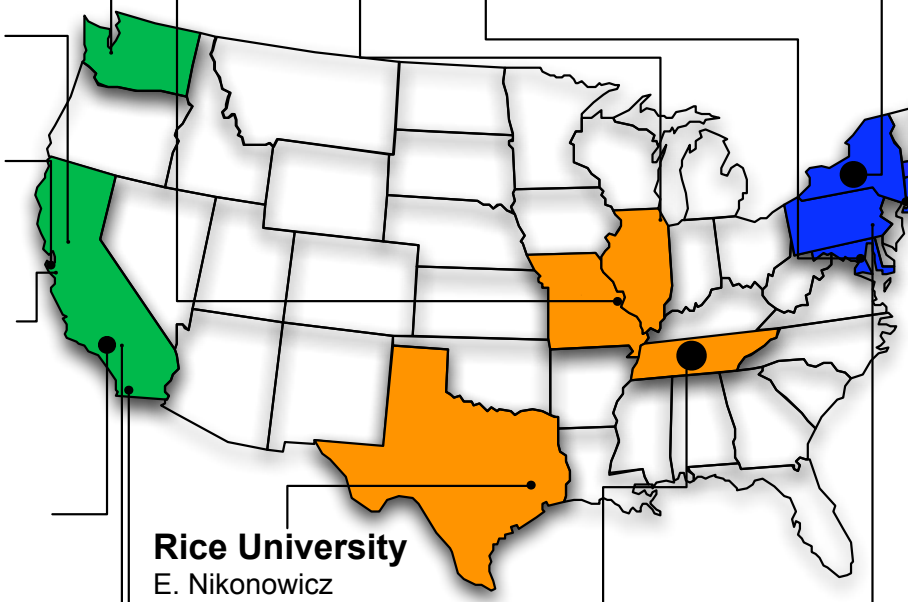
10 GB / sim

20 CPU-years

3 months (wall-clock)

ijstokes@hkl.hms.harvard.edu

SBGrid Consortium



Washington U. School of Med.
T. Ellenberger
D. Fremont

U. Washington
T. Gonen

UC Davis
H. Stahlberg

UCSF
JJ Miranda
Y. Cheng

Stanford
A. Brunger
K. Garcia
T. Jardetzky

CalTech
P. Bjorkman
W. Clemons
G. Jensen
D. Rees

WesternU
M. Swairjo

UCSD
T. Nakagawa
H. Viadiu

Rosalind Franklin
D. Harrison

NIH
M. Mayer

U. Maryland
E. Toth

Rice University
E. Nikonowicz
Y. Shamoo
Y.J. Tao

Vanderbilt
Center for Structural Biology
W. Chazin
B. Eichman
M. Egli
B. Lacy
C. Sanders
B. Spiller
M. Stone
M. Waterman

Thomas Jefferson
J. Williams

Cornell U.
R. Cerione
B. Crane
S. Ealick
M. Jin
A. Ke

NE-CAT
R. Oswald
C. Parrish
H. Sondermann

UMass Medical
W. Royer

Brandeis U.
N. Grigorieff

Tufts U.
K. Heldwein

Columbia U.
Q. Fan

Rockefeller U.
R. MacKinnon

Yale U.
T. Boggon
D. Braddock
Y. Ha
E. Lolis
K. Reinisch
J. Schlessinger
F. Sigworth
F. Zhou

Harvard and Affiliates
N. Beglova
S. Blacklow
B. Chen
J. Chou
J. Clardy
M. Eck
B. Furie
R. Gaudet
M. Grant
S.C. Harrison
J. Hogle
D. Jeruzalmi
D. Kahne
T. Kirchhausen
A. Leschziner
K. Miller
A. Rao
T. Rapoport
M. Samso
P. Sliz
T. Springer
G. Verdine
G. Wagner
L. Walensky
S.Walker
T.Walz
J. Wang
S. Wong

Not Pictured:
University of Toronto: L. Howell, E. Pai, F. Sicheri; NHRI (Taiwan): G. Liou; Trinity College, Dublin: Amir Khan

ijstokes@hkl.hms.harvard.edu

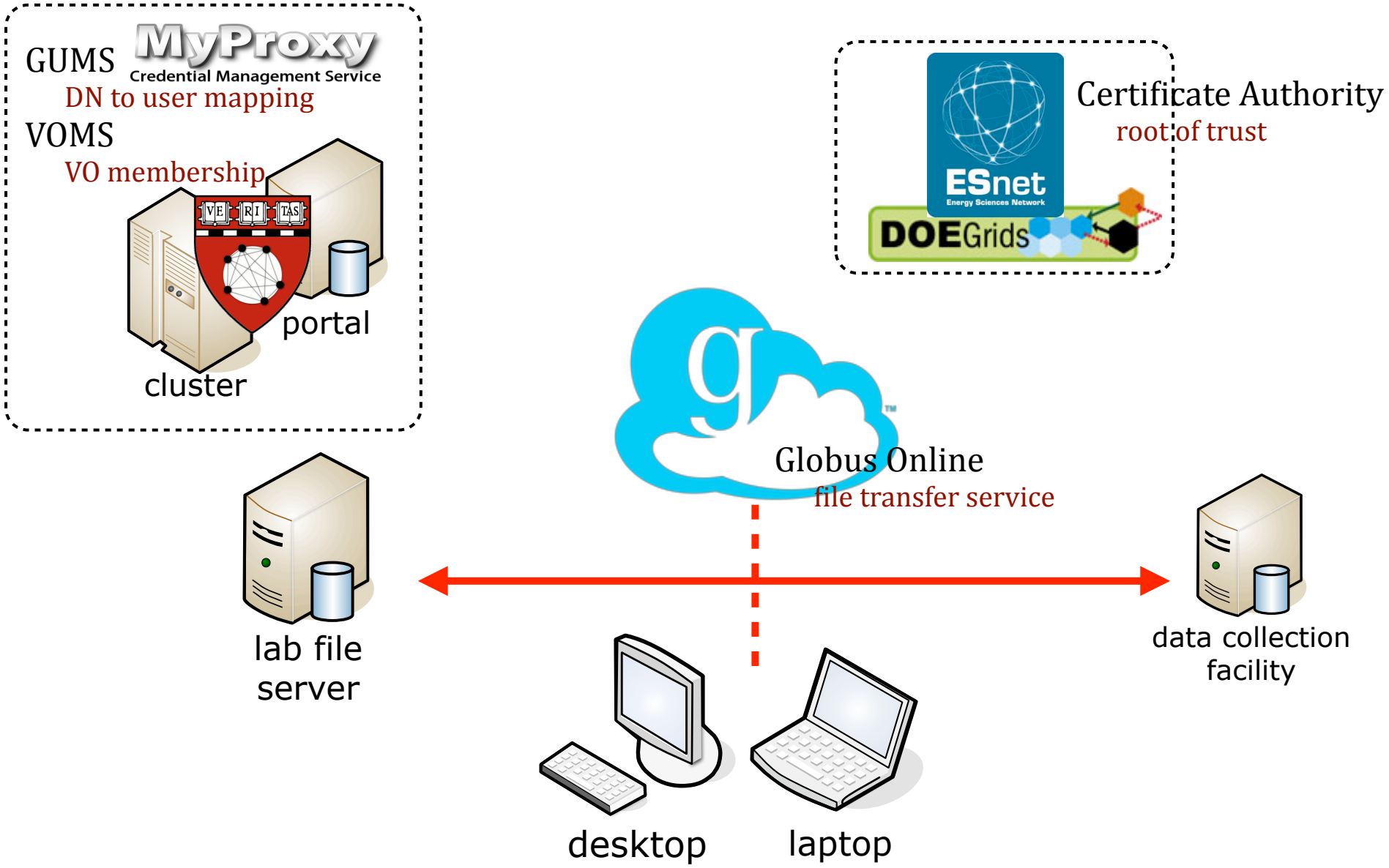
Boston Life Sciences Hub

- Biomedical researchers
- Government agencies
- Life sciences
- Universities
- Hospitals

ijstokes@hkl.hms.harvard.edu

Data Access

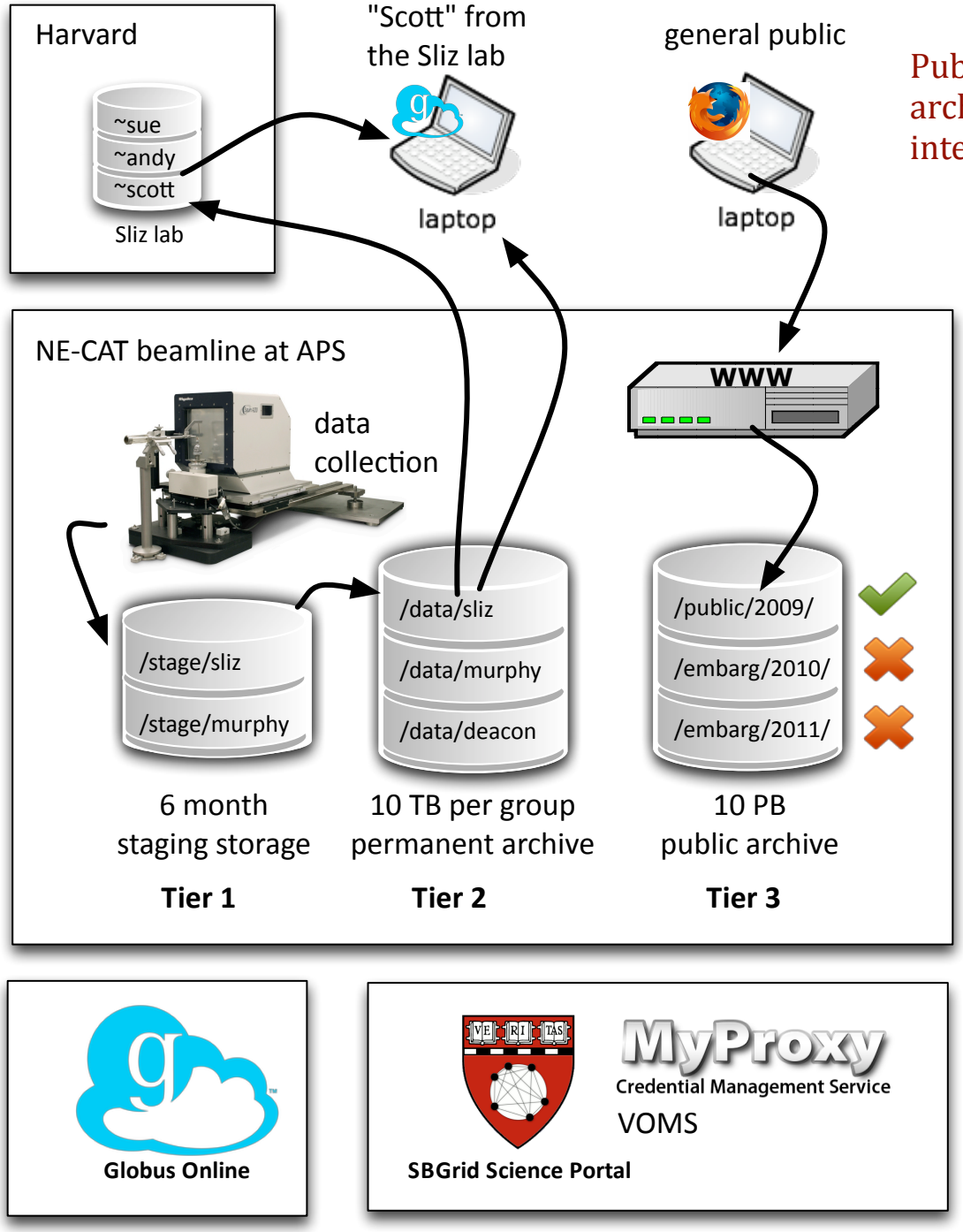
Globus Online: High Performance Reliable 3rd Party File Transfer



User can directly access lab or facility data from laptop

Local accounts within lab infrastructure

Public access available to archived data through web interface



Shared (lab level) accounts at facility

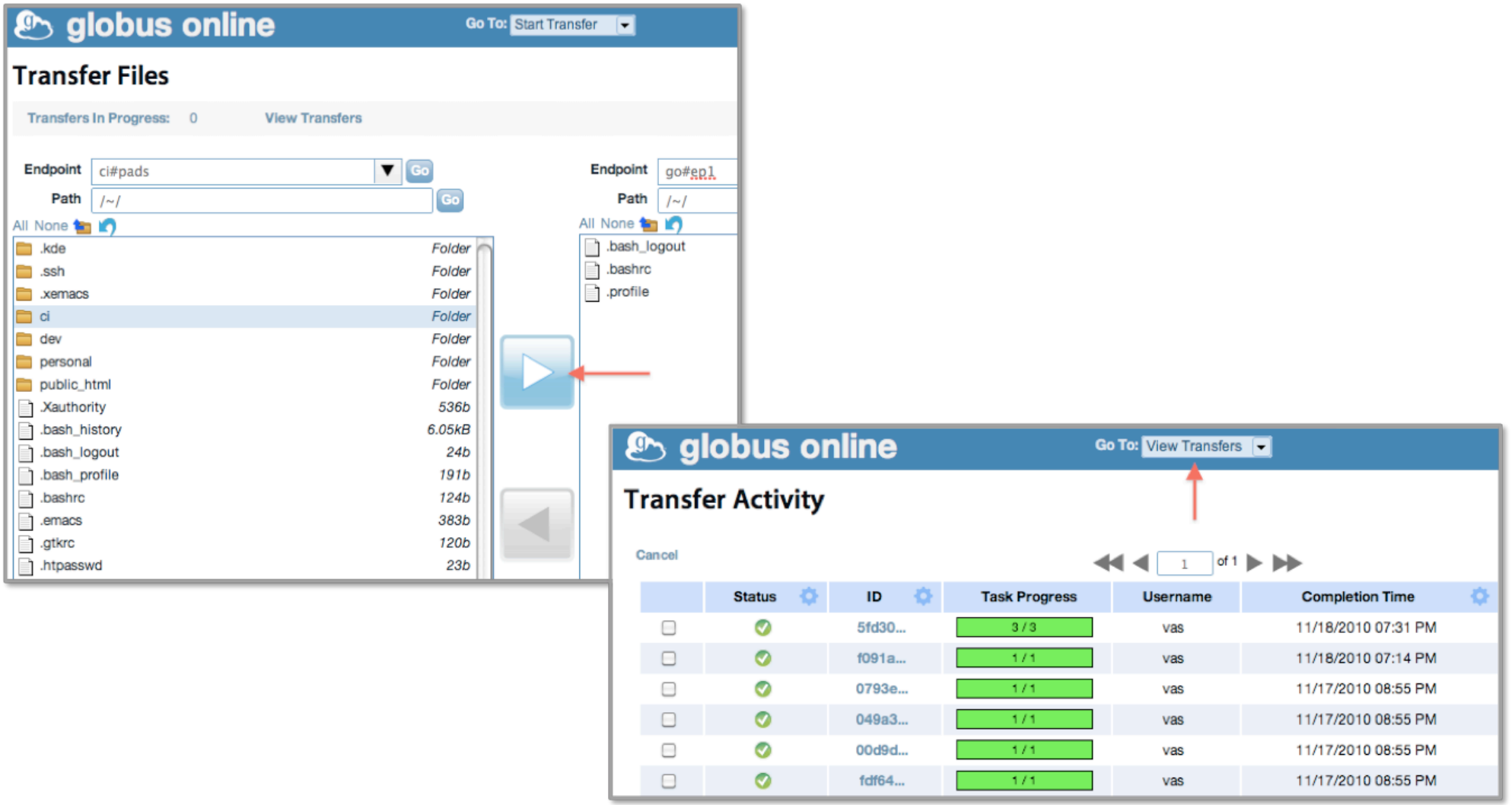
Tiered storage

Embargo policy to hold deposited data for agreed time

VO management

Architecture

- SBGrid
  - manages all user account creation and credential mgmt
  - hosts MyProxy, VOMS, GridFTP, and user interfaces
- Facility
  - knows about lab groups
    - e.g. “Harrison”, “Sliz”
  - delegates knowledge of group membership to SBGrid VOMS
    - facility can poll VOMS for list of current members
  - uses X.509 for user identification
  - deploys GridFTP server
- Lab group
  - designates group manager that adds/removes individuals
  - deploys GridFTP server or Globus Connect client
- Individual
  - username/password to access facility and lab storage
  - Globus Connect for personal GridFTP server to laptop
  - Globus Online web interface to “drive” transfers


globus online
Go To: Start Transfer
Transfer Files
Transfers In Progress: 0
View Transfers
Endpoint ci#pads
Go
Path /~/
Go
All None
.kde Folder
.ssh Folder
.xemacs Folder
ci Folder
dev Folder
personal Folder
public_html Folder
.Xauthority 536b
.bash_history 6.05kB
.bash_logout 24b
.bash_profile 191b
.bashrc 124b
.emacs 383b
.gtkrc 120b
.htpasswd 23b
Endpoint go#ep1
Path /~/
All None
.bash_logout
.bashrc
.profile
globus online
Go To: View Transfers
Transfer Activity
Cancel
1 of 1
Status
ID
Task Progress
Username
Completion Time
5fd30... 3 / 3 vas 11/18/2010 07:31 PM
f091a... 1 / 1 vas 11/18/2010 07:14 PM
0793e... 1 / 1 vas 11/17/2010 08:55 PM
049a3... 1 / 1 vas 11/17/2010 08:55 PM
00d9d... 1 / 1 vas 11/17/2010 08:55 PM
fdf64... 1 / 1 vas 11/17/2010 08:55 PM

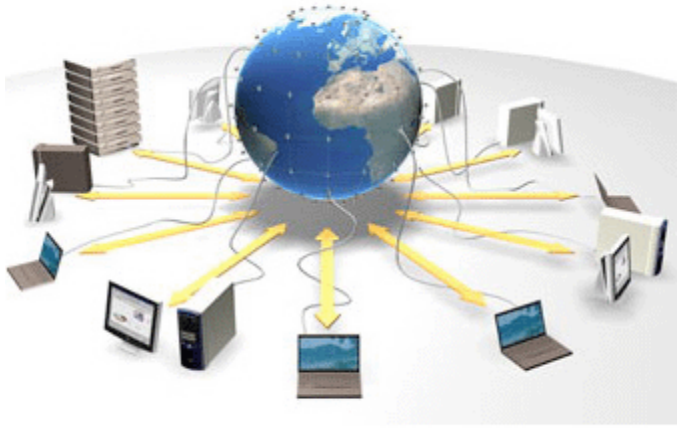
Objective

- Easy to use high performance data mgmt environment
- Fast and reliable file transfer
  - facility-to-lab, facility-to-individual, lab-to-individual
- Reduced administrative overhead
- Better data curation
- Release data to public after embargo period

Challenges

- Access control
  - visibility
  - policies
- Provenance
  - data origin
  - history
- Meta-data
  - attributes
  - searching

Grid Computing Today



Capabilities

- Server-to-server interaction
- Federated aggregation of CPU power
- Predictable data patterns
- Command-line access from specialized servers
- Web-based access through “Science Gateways”
  - pre-defined computing workflows
  - e.g. SBGrid Science Portal

Weaknesses

- Federated user identity system
  - X.509 digital certificates and associated apparatus
- Ease of use for end users
- Data management tools
- Support for collaborations

Opportunities

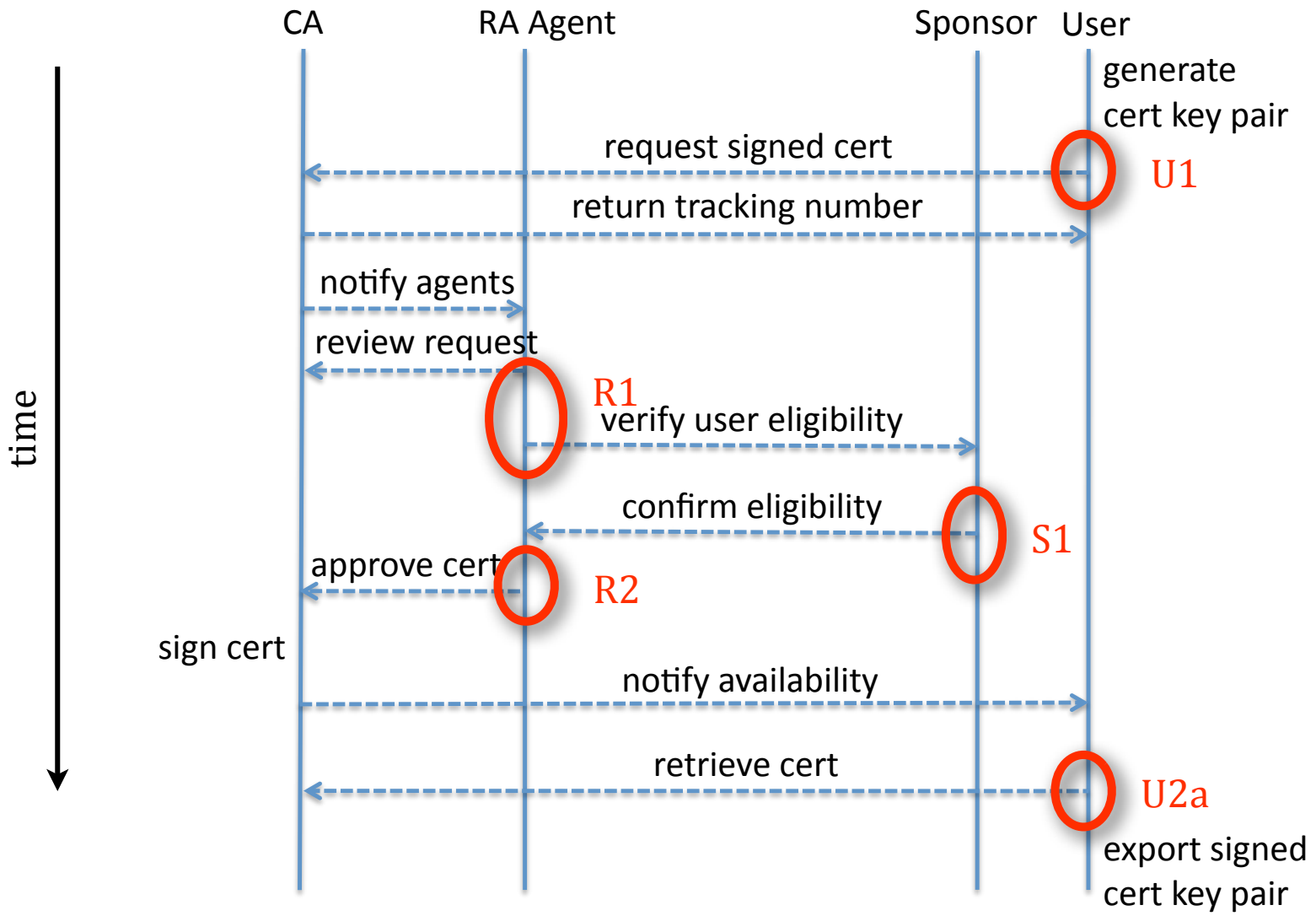
- "Last mile" challenge
  - to the desktop
  - to the laptop
- Unified identity management
  - centralized set of credentials for each person
- Empower collaborations to self-manage
- Shift of focus from "compute" to "data"
  - for users
  - for facilities where data is the main challenge

User Credentials

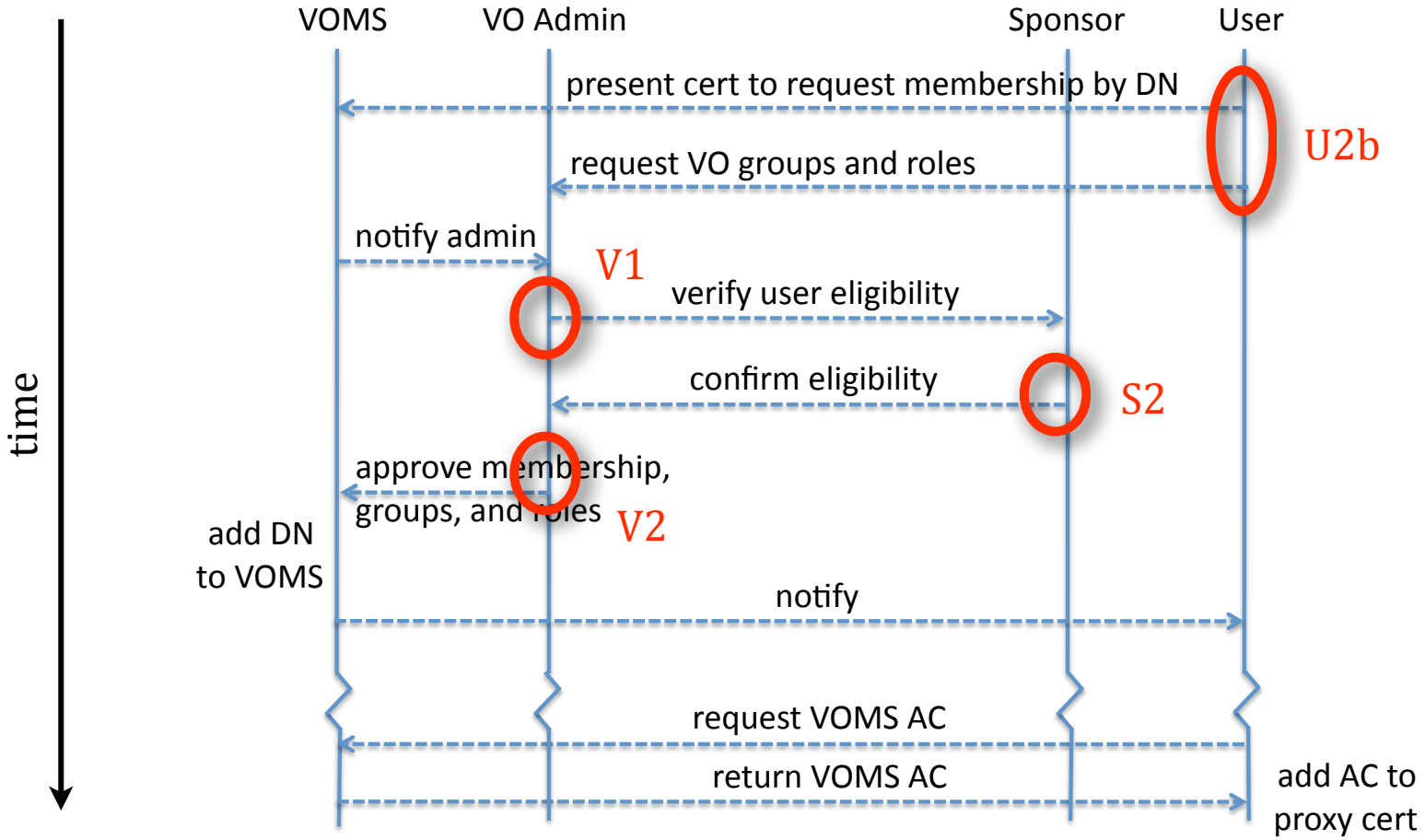
X.509 Digital Certificates

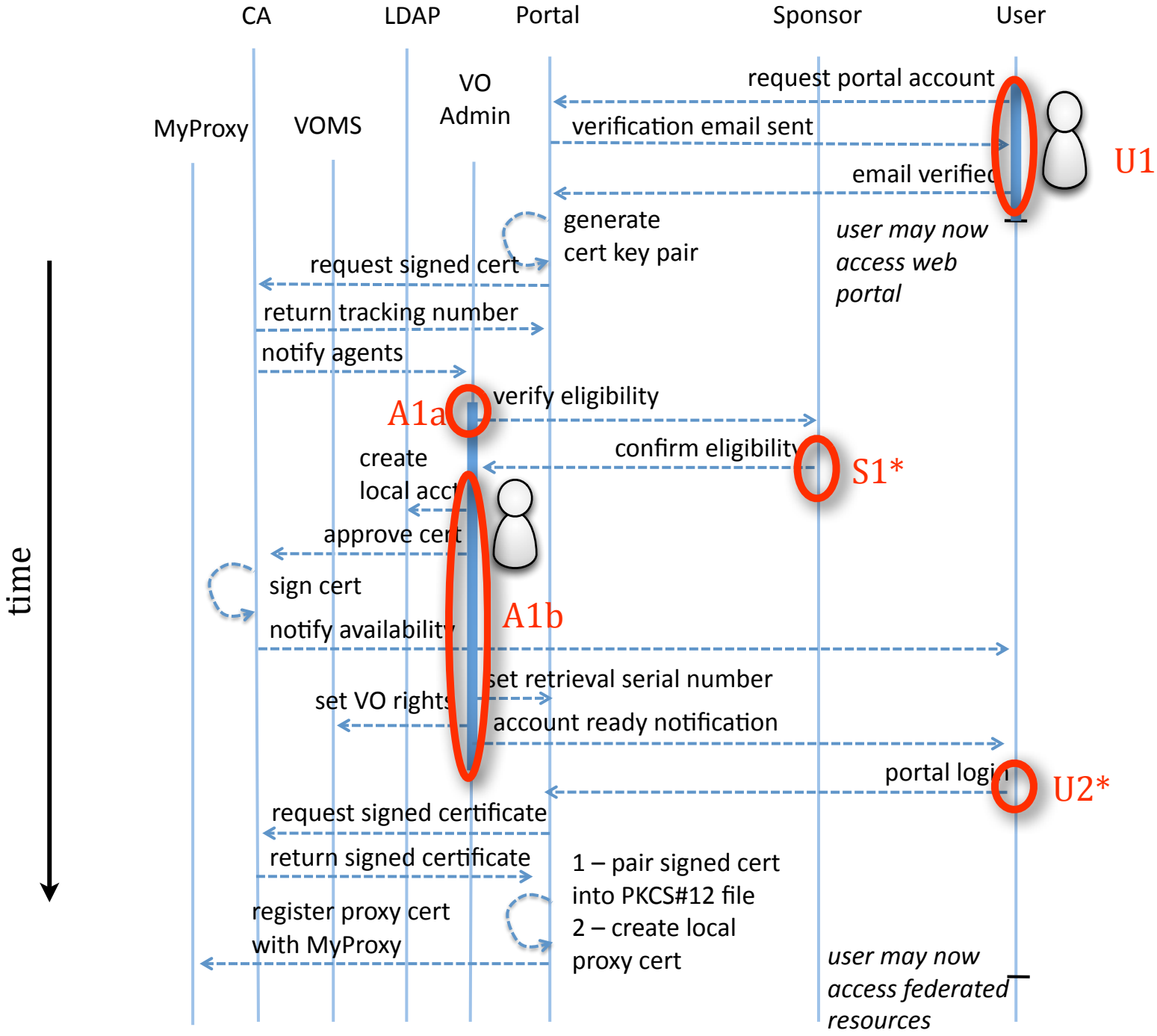
- Analogy to a passport:
  - Application form
  - Sponsor's attestation
  - Consular services
    - verification of application, sponsor, and accompanying identification and eligibility documents
  - Passport issuing office
- Portable, digital passport
  - fixed and secure user identifiers
    - name, email, home institution
  - signed by widely trusted issuer
  - time limited
  - ISO standard

Addressing Certificate Problems



VO (Group) Membership Registration




CA
LDAP
Portal
Sponsor
User
MyProxy
VOMS
VO Admin
request portal account
verification email sent
email verified
U1
generate cert key pair
user may now access web portal
request signed cert
return tracking number
notify agents
A1a
verify eligibility
confirm eligibility
S1*
create local acct
approve cert
sign cert
notify availability
A1b
set retrieval serial number
set VO rights
account ready notification
portal login
U2*
request signed certificate
return signed certificate
1 – pair signed cert into PKCS#12 file
2 – create local proxy cert
register proxy cert with MyProxy
user may now access federated resources
time

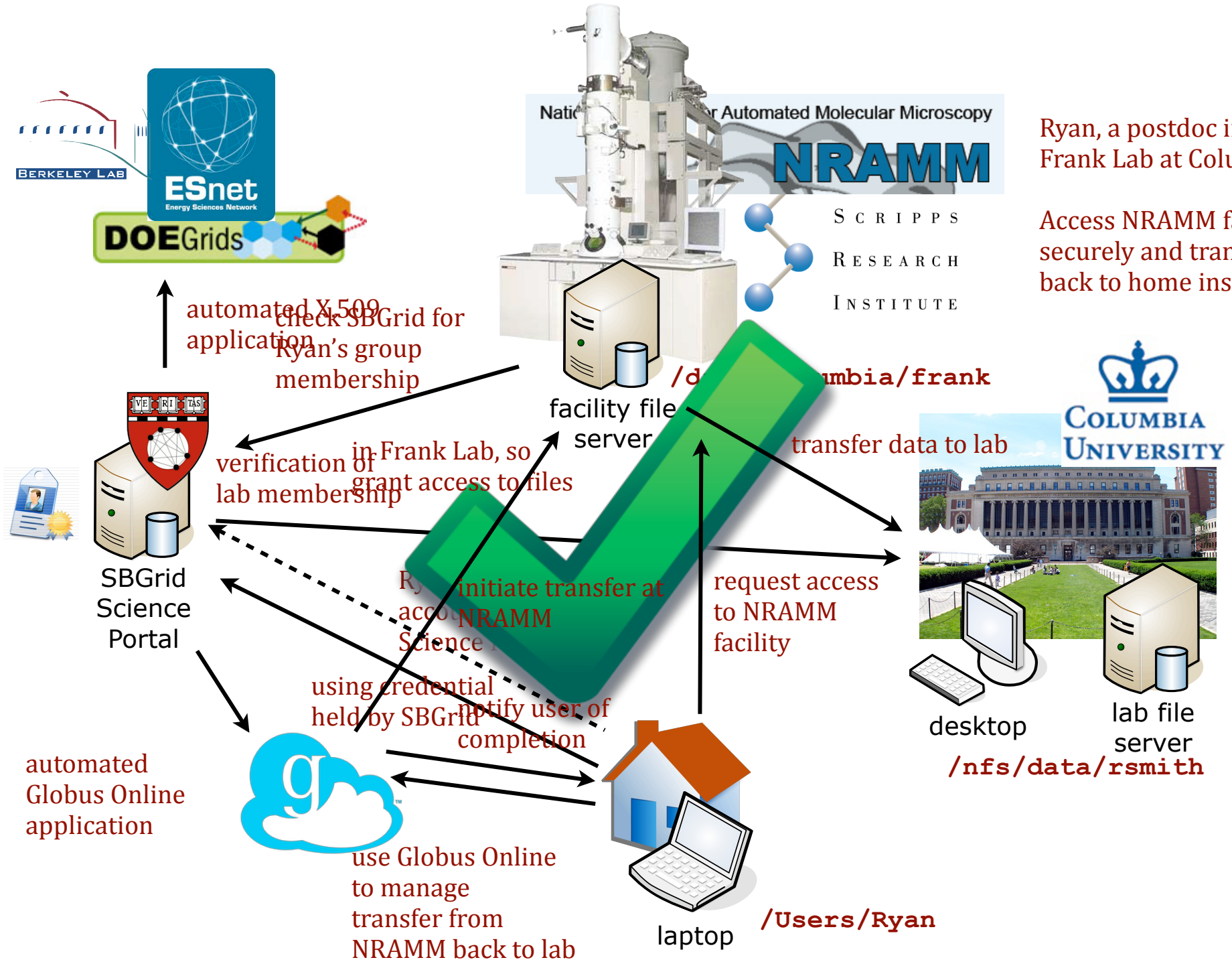
Process and Design Improvements

- Single web-form application
  - includes e-mail verificationn
- Centralized and connected credential management
  - FreeIPA LDAP - user directory and credential store
  - VOMS - lab, institution, and collaboration affiliations
  - MyProxy - X.509 credential store
- Overlap administrative roles
  - system admin
  - registration agent for certificate authority (approve X.509 request)
  - VO administrator to register group affiliations
- Automation

“Last Mile” and Ease of Use

Grid computing at your desk

- Platforms
  - Windows
  - OS X
- Web interfaces
  - SBGrid Science Portal
- One-stop-shop for account management
  - Centralized services
  - Automated processes
  - Administrator intervention
  - Support

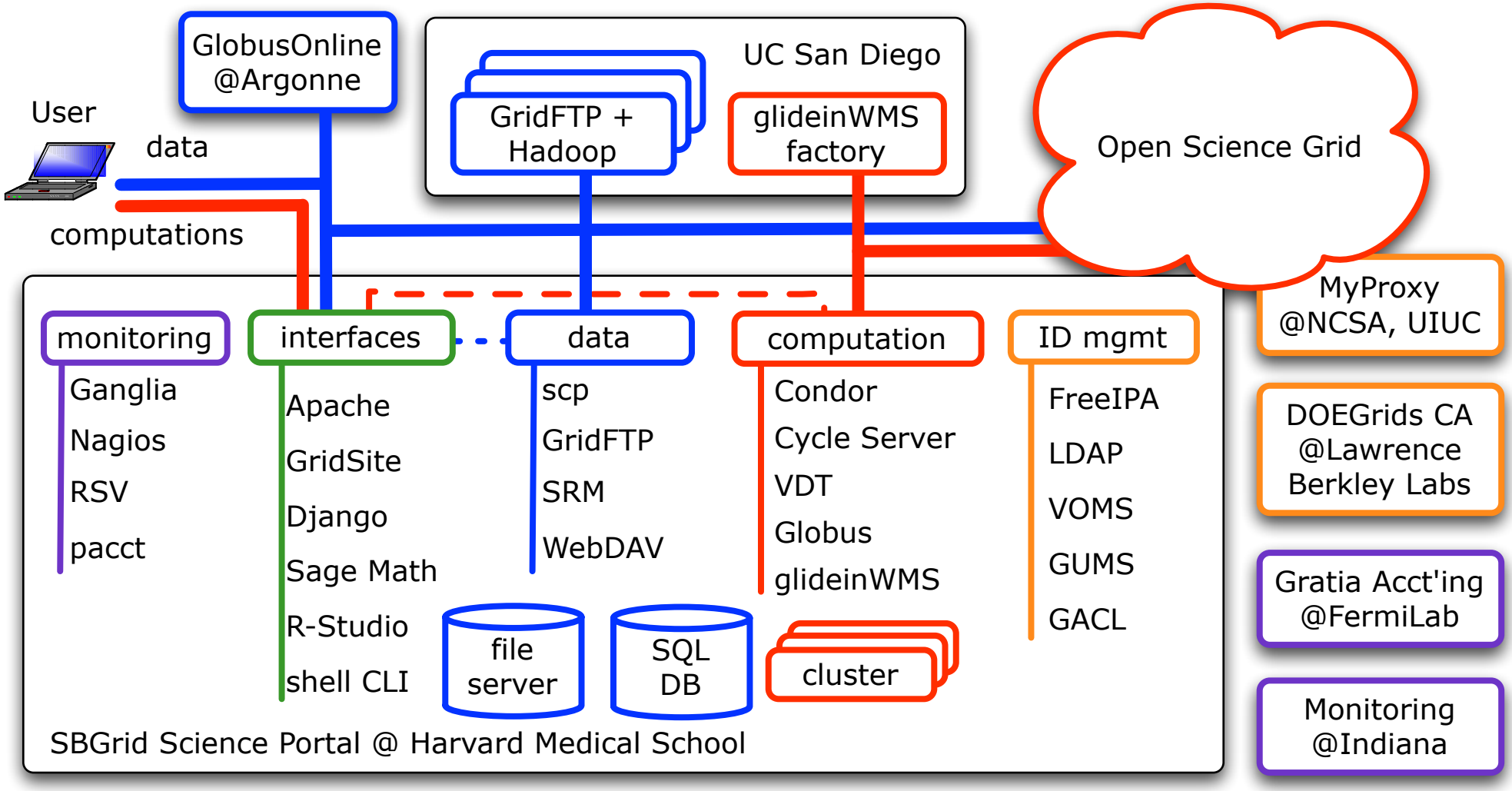


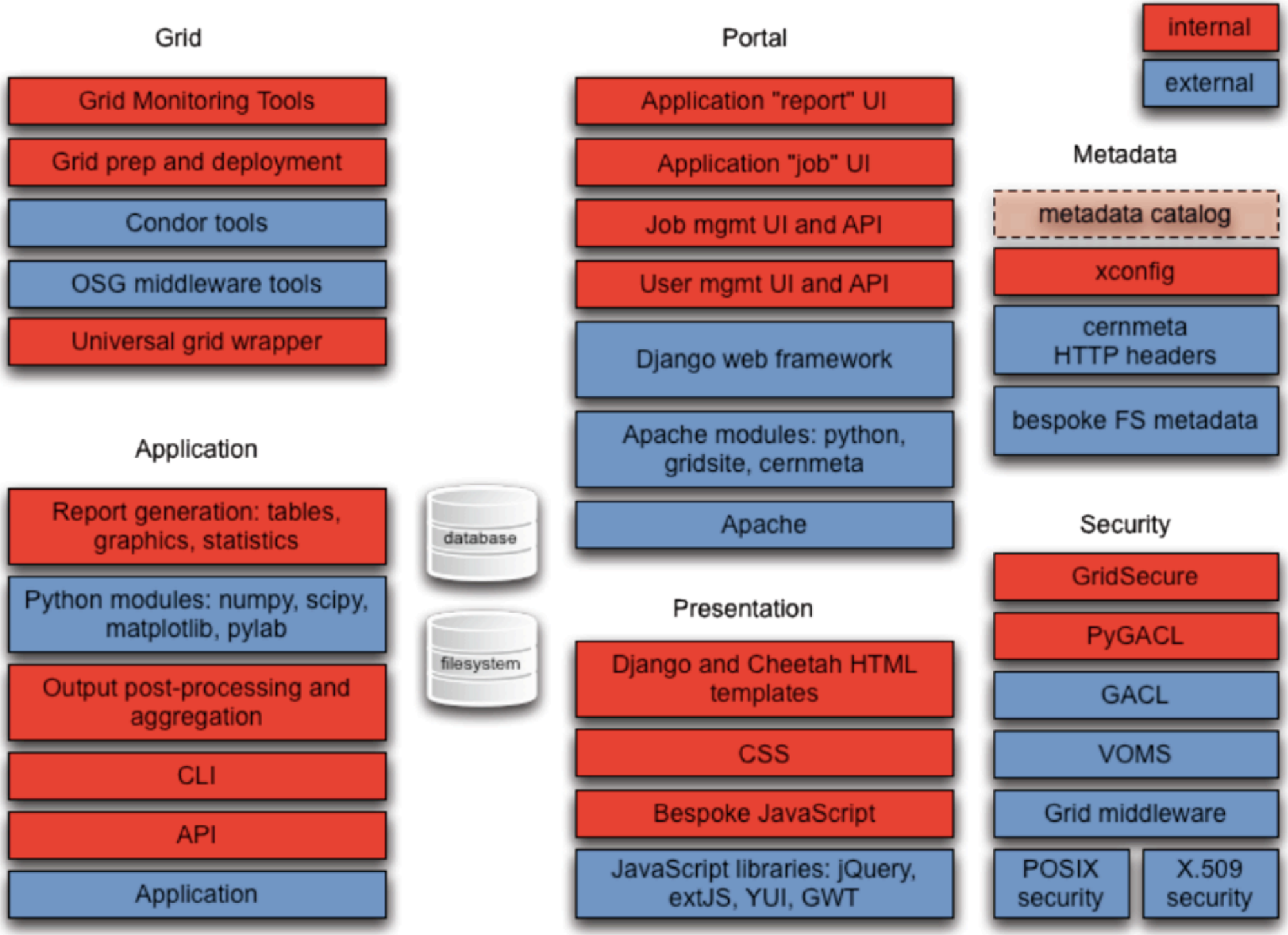
Ryan, a postdoc in the Frank Lab at Columbia

Access NRAMM facilities securely and transfer data back to home institute

Architecture Diagrams

Service Architecture





ijstokes@hkl.hms.harvard.edu

Acknowledgements

- Piotr Sliz
- SBGrid Science Portal
  - Daniel O'Donovan, Meghan Porter-Mahoney
- SBGrid System Administrators
  - Ian Levesque, Peter Doherty, Steve Jahl
- Facility Collaborators
  - Frank Murphy (NE-CAT/APS)
  - Ashley Deacon (JCSG/SLAC)
- Globus Online Team
  - Steve Tueke, Ian Foster, Rachana Ananthakrishnan, Raj Kettimuthu
- Ruth Pordes
  - Director of OSG, for championing SBGrid