Working Meeting at FermiLab

University of California Grid (UC Grid)

January 19-20, 2010

Fermi Lab, Chicago, IL

LBNL

Krishna Muriki, Jackie Scoggins, Tammy Welcome,
Shreyas Chola, Doug Olson, Gary Jung, Greg Kurtzer
Lawrence Berkeley National Laboratory

William Strossman, Russ Harvey, Bob Grant
University of California, Riverside

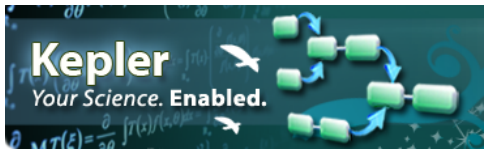
Harry Mangalam, Joseph Farran, Frank Wessel
University of California, Irvine

Kejian Jin, Shao-Ching Huang, TV Singh, Sveta Mazurkova, Bill Labate, Qiyang Hu, Prakashan Korambath, Seonah Kim, Scott Johnson, Puneet Souda

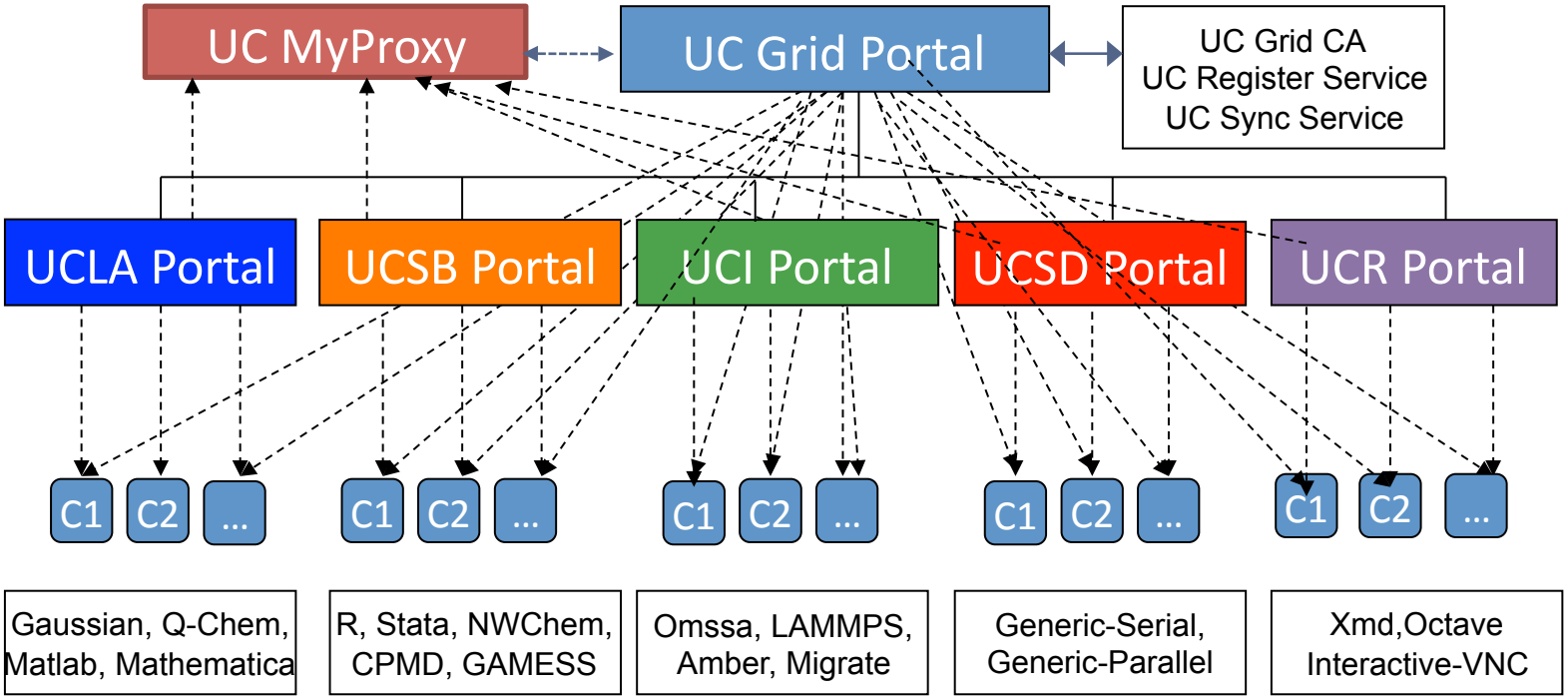
Paul Weakliem, Arlene Allen
University of California, Santa Barbara

David Walker, Bill Broadley, Russ Hobby
University of California, Davis

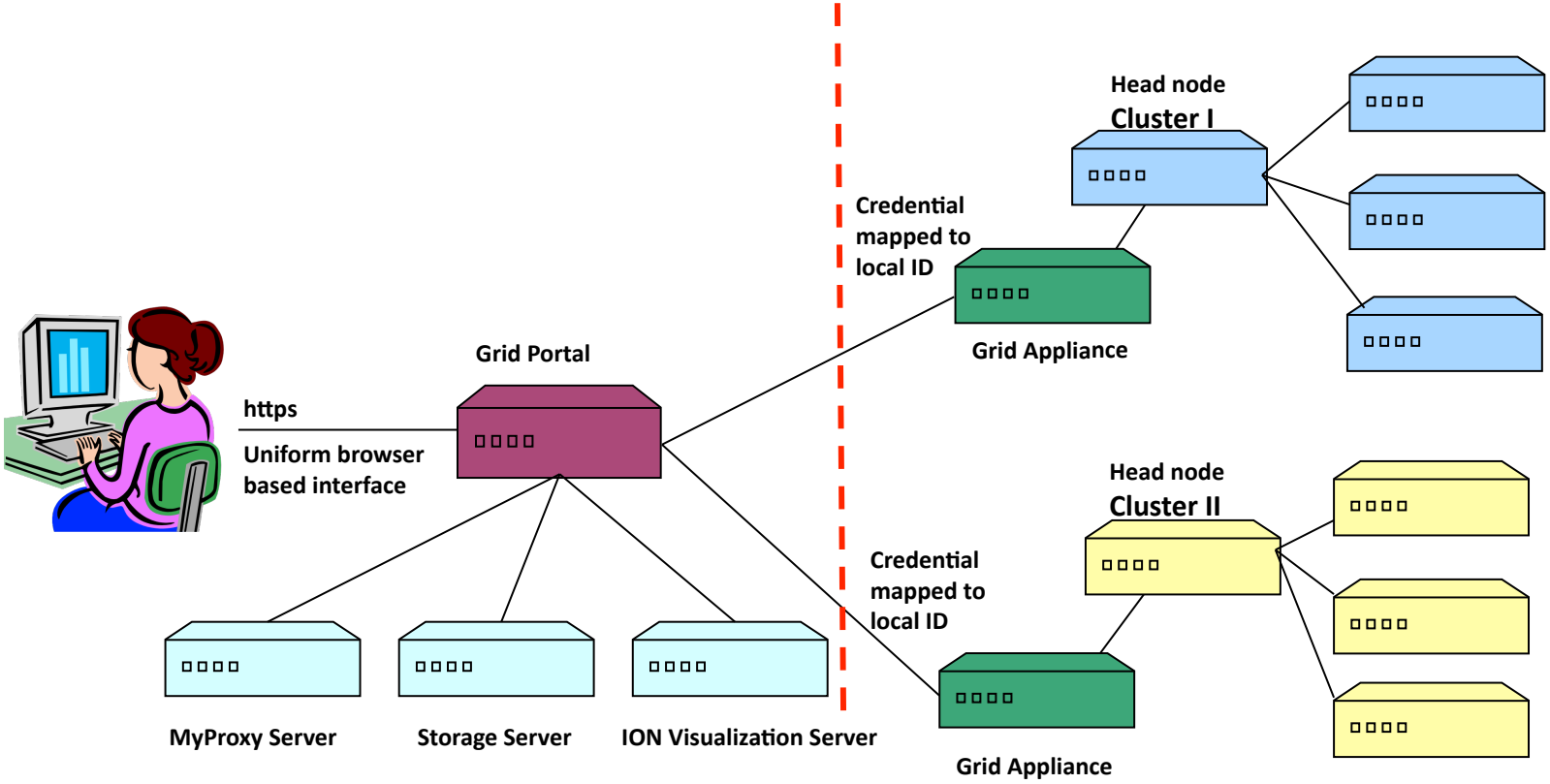
Shava Smallen, Jianwu Wang, Ilkay Altinas,
Richard Moore, David Hutches
San Diego Super Computing Center



UC Grid Architecture



Single Campus Architecture



The UC Grid Portal

- Provides Register Service (SOA):
  - user portal account creation and uniqueness of user name in the UC Grid, automatic CA sign, MyProxy push.

- Provides Sync Service (SOA) between campus portal and UC Grid portal:
  - Any change of an user account at a campus Grid Portal is automatically propagated to the UC Grid Portal.
  - Any change of new resources at a campus Grid Portal will also be reflected at UC Grid Portal

- From the UC Grid Portal, users can:
  - work with clusters from all the campuses those users are authorized to use.
  - transfer files across the campuses.
  - Use pooled resources across the campuses.

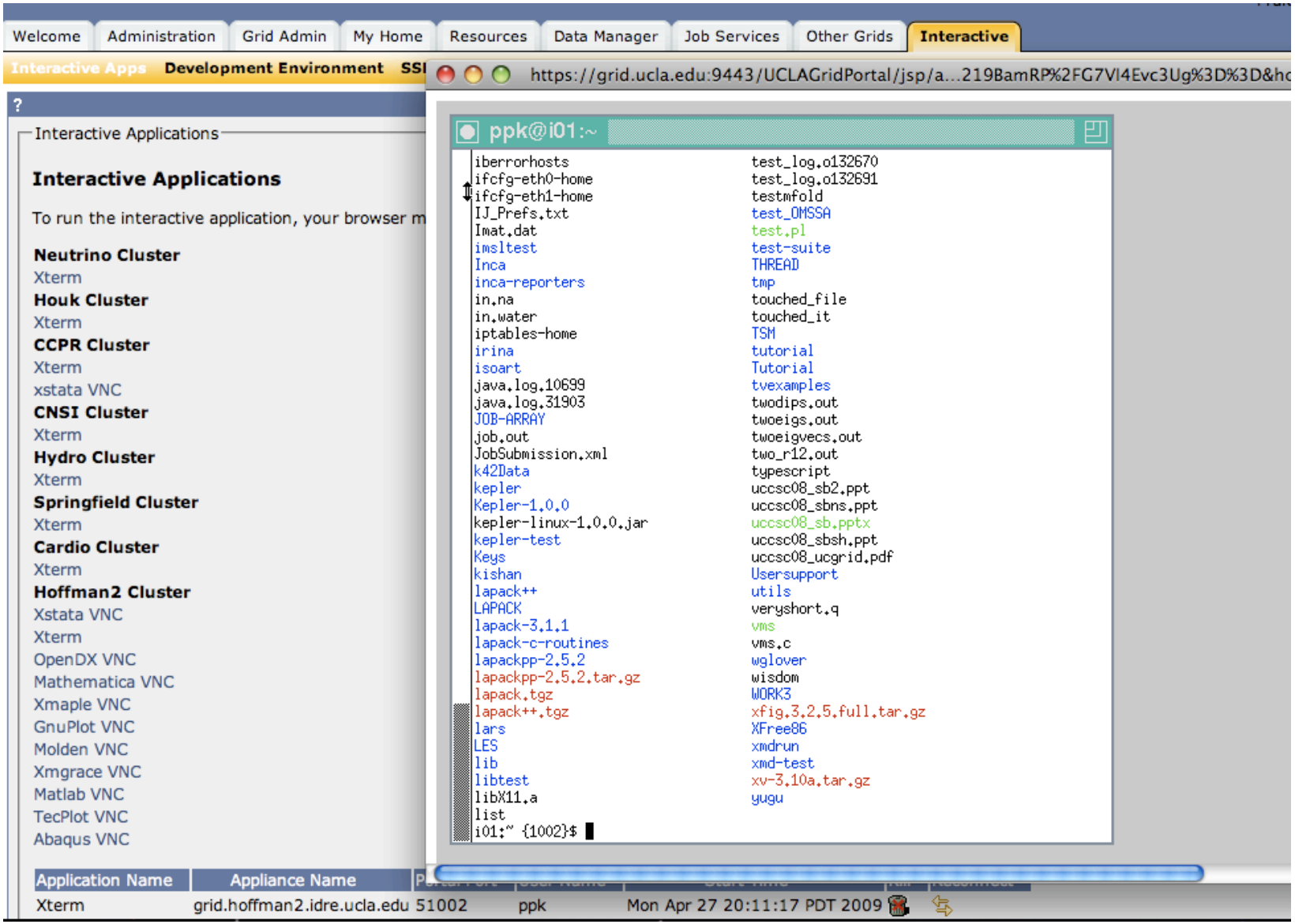
Campus Grid Portal

- Features:
  - Resource Discovery
    - Cluster load, usage, queue information, job details
  - Data Manager
    - File management such as create, remove, rename, edit, transfer etc.
  - Job Services
    - Submit job as a cluster user
    - Submit job as a pool user
  - Interactive X window for applications
    - Matlab, Mathematica, Maple, Xstata etc.
  - Access to Teragrid

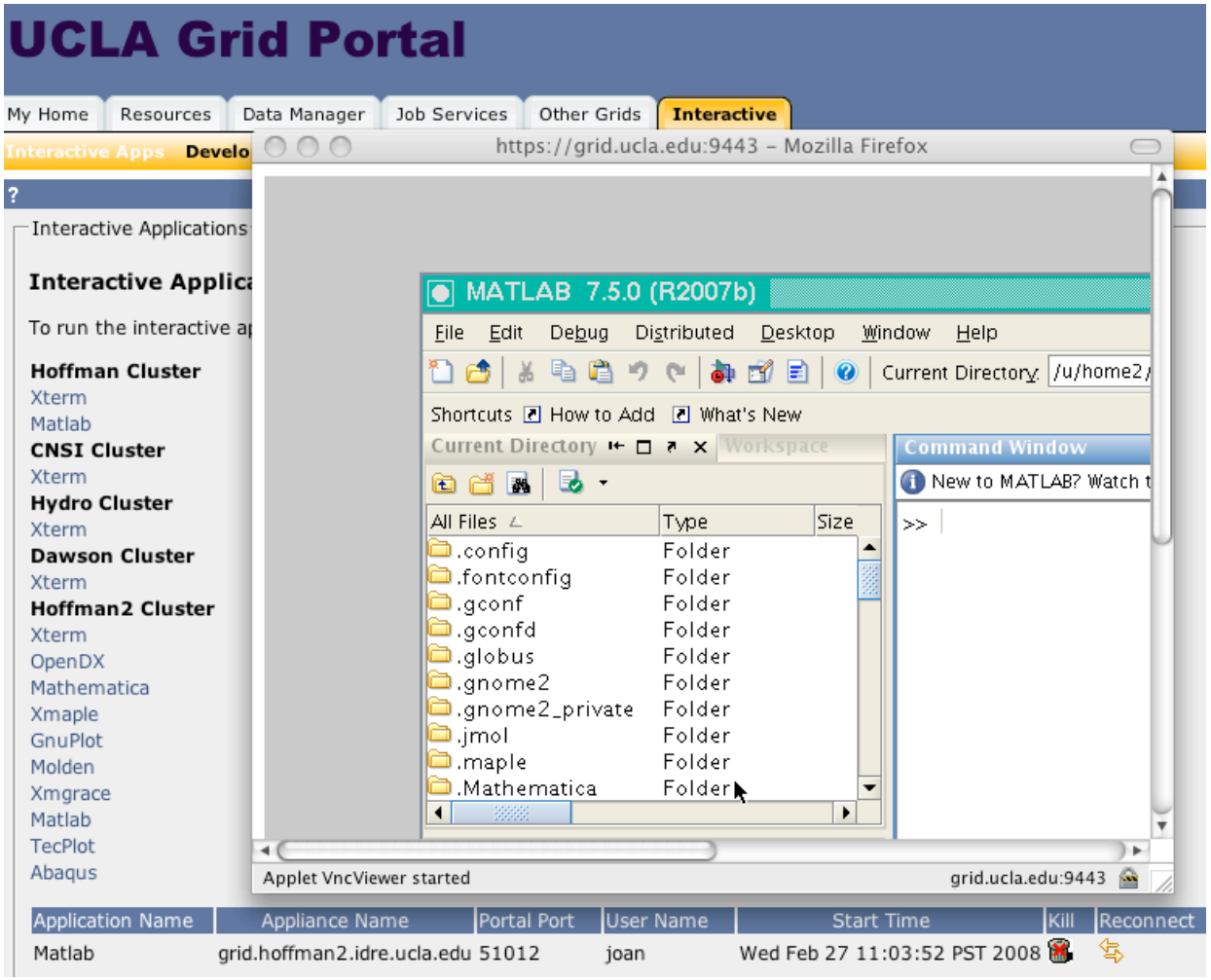
Job Submission

- General Job submission
  - Any user created executable
- Application job submission
  - Predefined executable
    - As Cluster user
    - As Pool user
- Pool user application job submission
  - UGP Chooses the target cluster for the application job.
  - Stages the input files to a guest user login id at the target cluster
  - Submits the job to run under a guest login id at the target cluster.
  - Local scheduler determines when those jobs are run.
  - UGP facilitates the transfer of the output files back to the Pool User.
  - Provide job monitoring and auditing

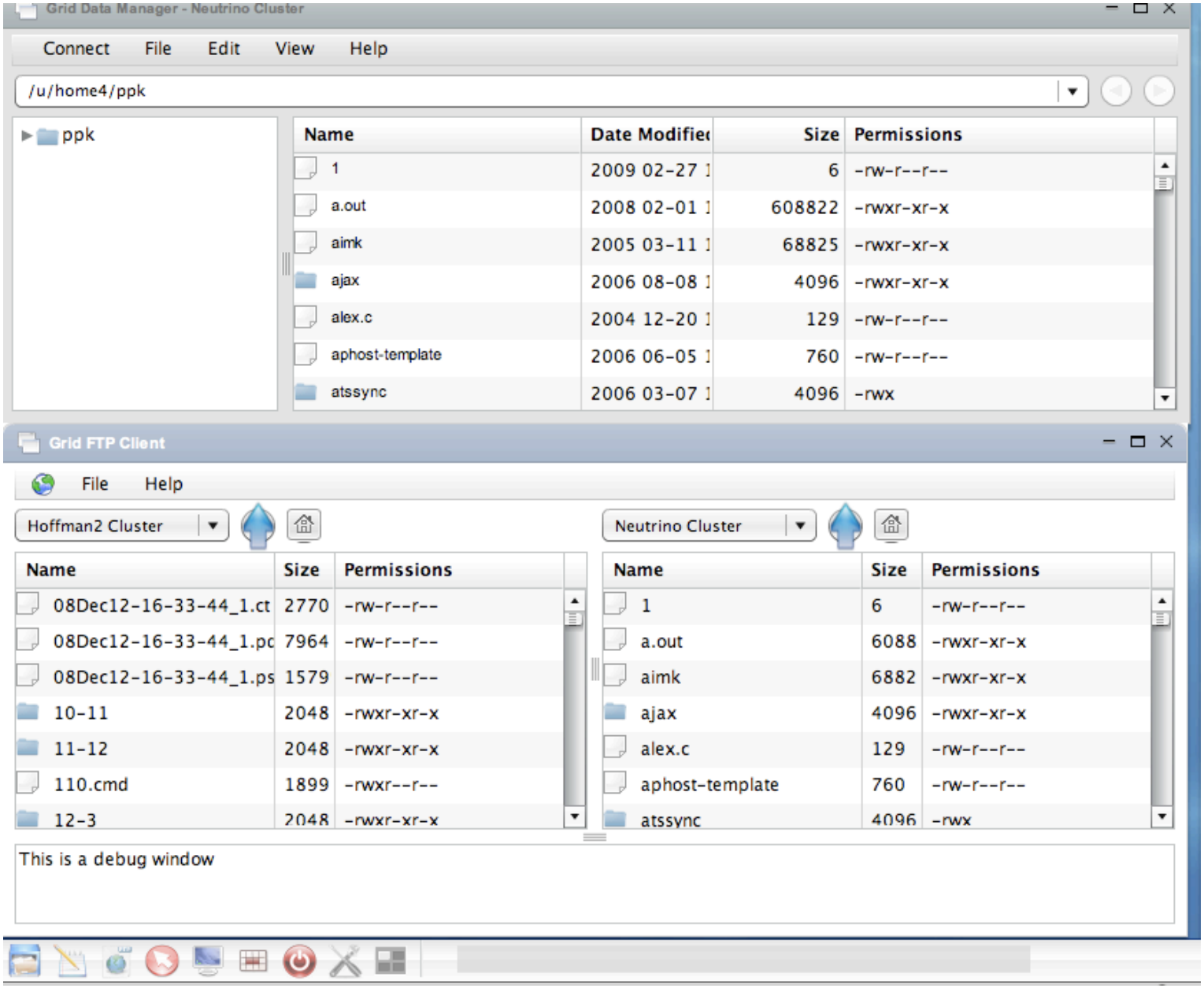
Interactive Xterm through VNC



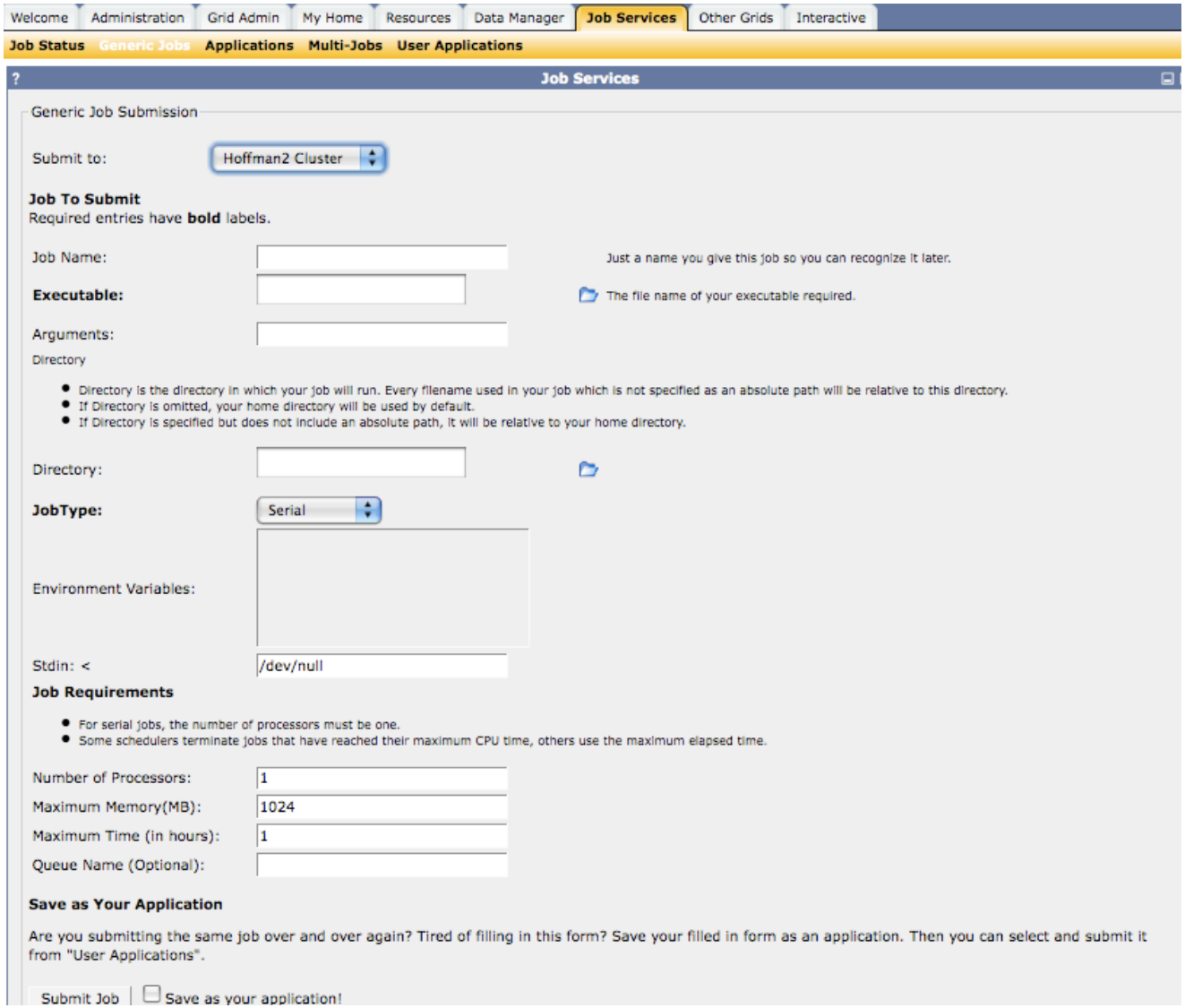
Interactive GUI Application through VNC



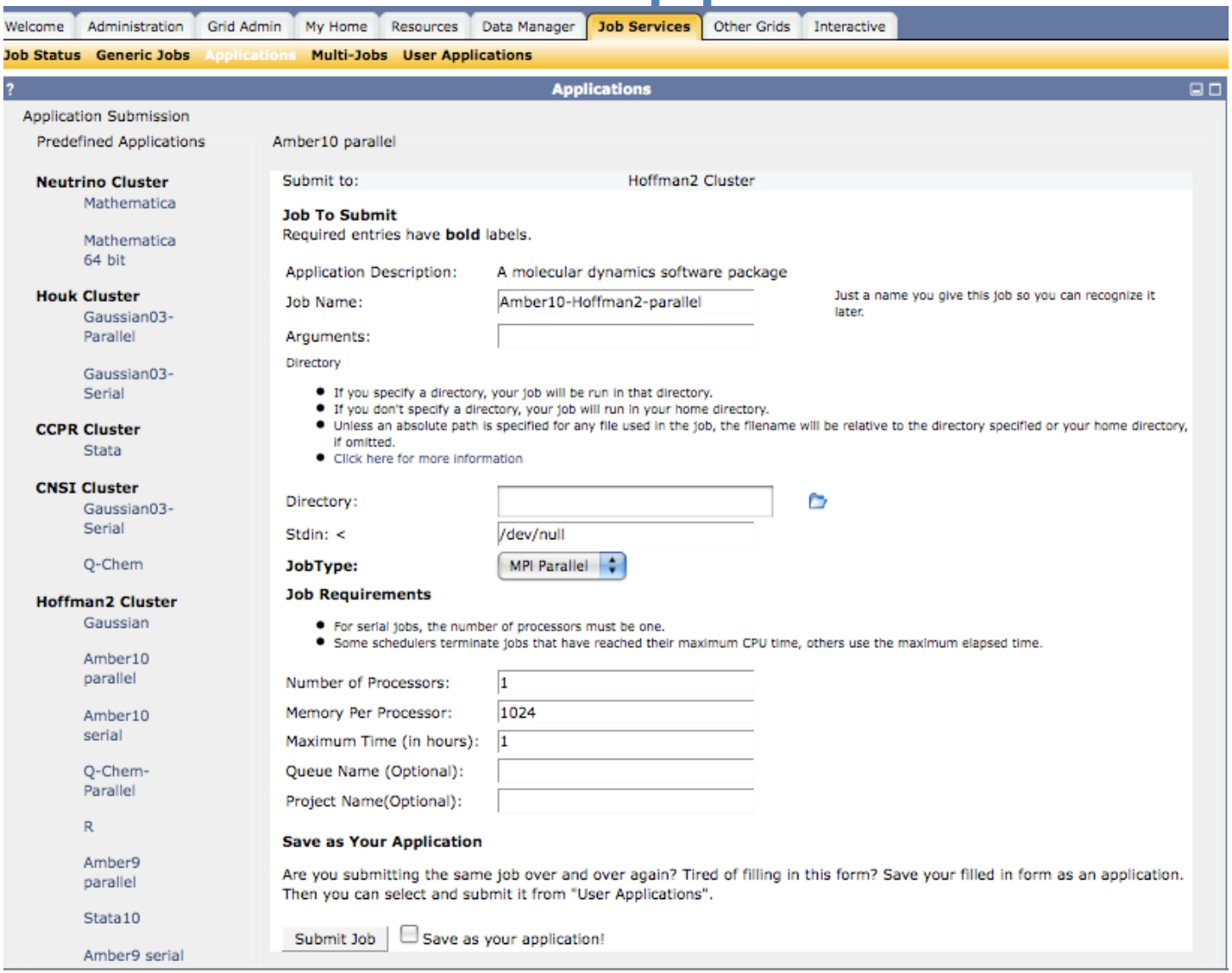
File Transfer



Generic Batch Job Submission Interface

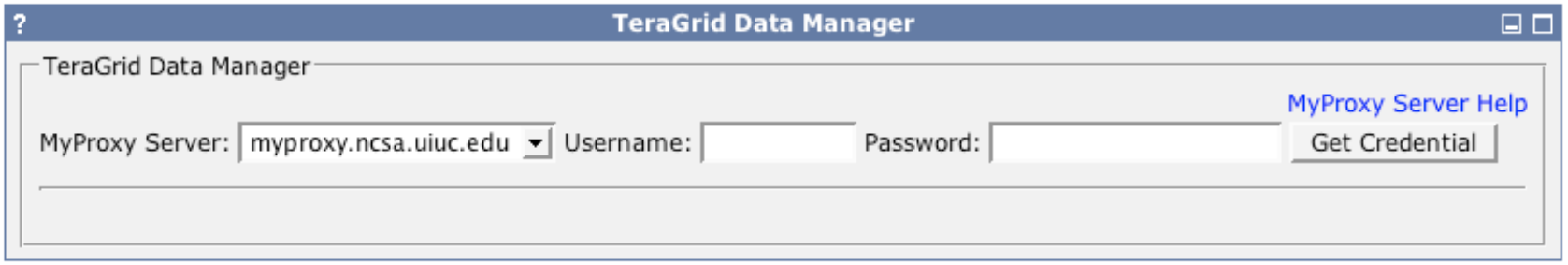


Predefined Application



Job Submission to Other Grid Resources

User interface to fill in username and password to retrieve short lived credential for submitting jobs to Teragrid



Proteomics Workflow Project

Mass Spectrometer

Binary Data/
Converter

Pre-Processed Data
(OMSSA/InsPect Format,
Commercial Application)

Grid Temp
Storage

Grid Portal

User
Generated
Query

OMSSA/InsPect
Run

NIH BLAST Library
(Downloaded to Cluster)

Meta-scheduler

Cluster A

Cluster B

Cluster C

Data Merge/Split
(based on job
type)

1. Single job or job array
2. Output collection

OMSSA Output
to user's desktop,
OMSSA Viewer

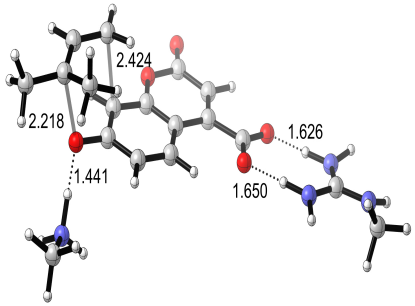
InsPect Output
to user's desktop,
Web Viewer

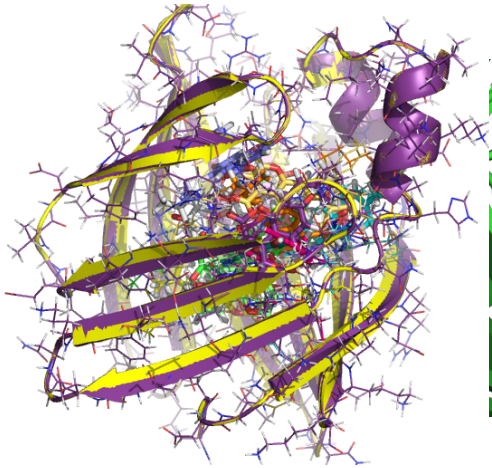
Protein Design Processes

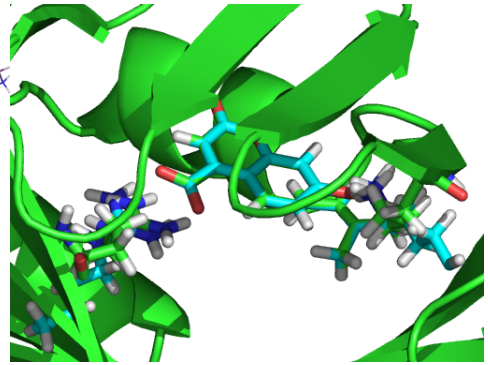
UCLA

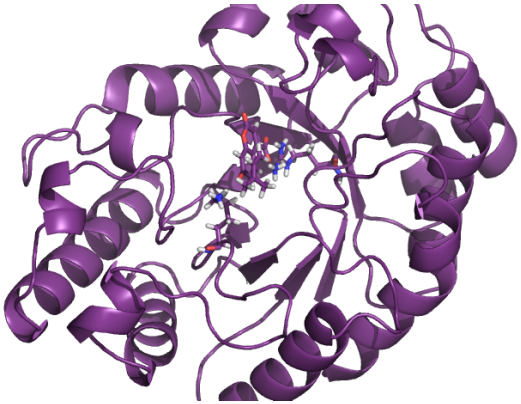
Inside-out protocol to **design novel enzyme catalysts**

Baker - Houk collaboration









Transition structure (QM theozyme)

Match theozyme geometry to protein scaffold *(RosettaMatch)*

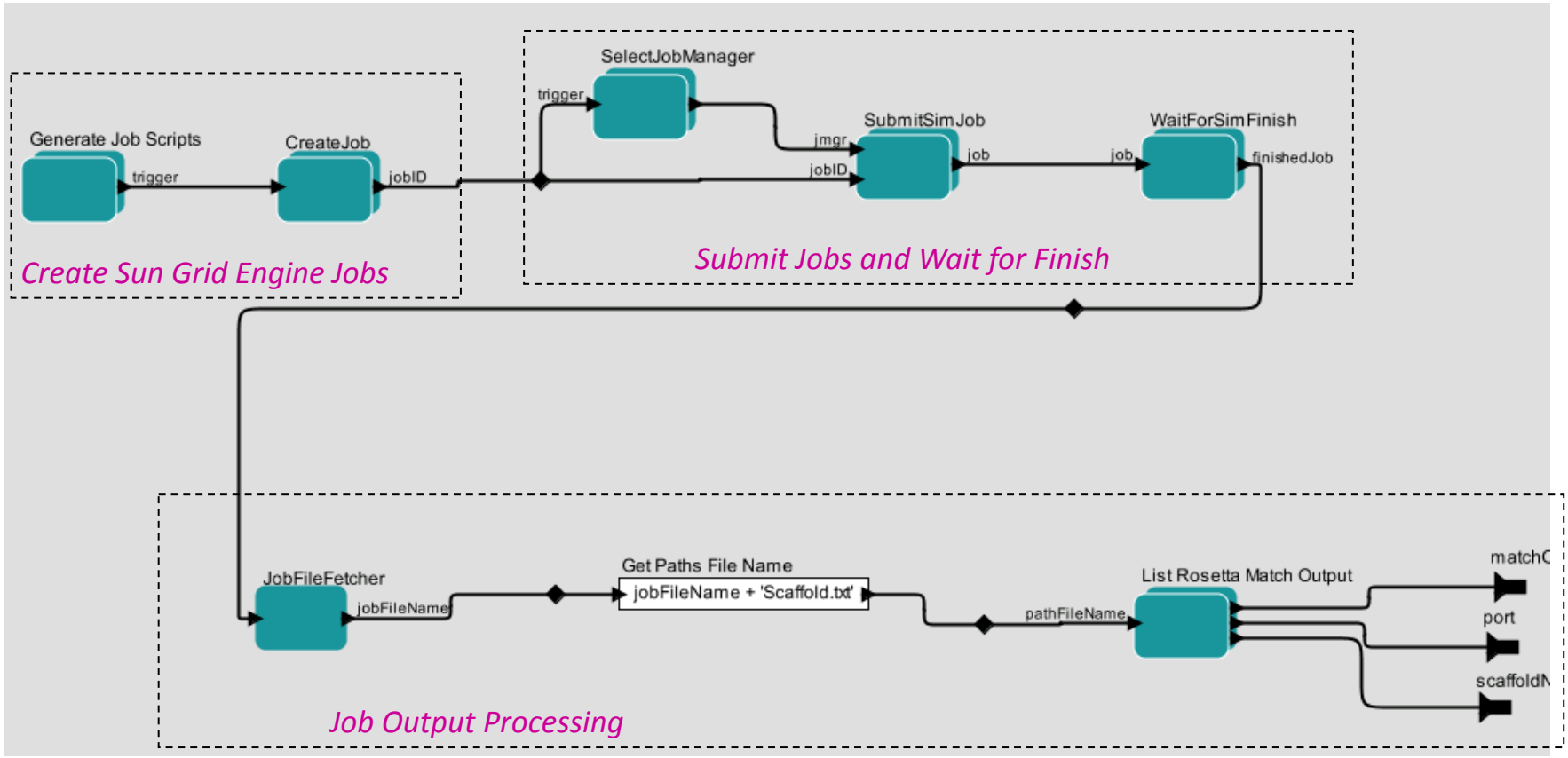
EDGE (Rank and Filter)

Redesign and repack Active site *(RosettaDesign)* & MD/EDGE

Enzyme Design Using Kepler Workflow

- RosettaMatch
  - Goal: Finds the enzyme Scaffolds for the input theozyme
  - Approximately 250 Jobs per submission – Each creates around 2000 outputs
  - Workflow Distribute the Jobs through Scheduler/Grid
- Edge
  - Goal: Filter out Match results that are out of user defined range
  - Least time consuming – Single node job
- RosettaDesign
  - Goal: Repacks and Optimize the side chains
  - Workflow Distribute 5 million jobs from Match run through Scheduler/Grid

Rosetta Match SGE Composite Actor



Command line instructions for two cluster workflow (Hoffman2 and Hydro)

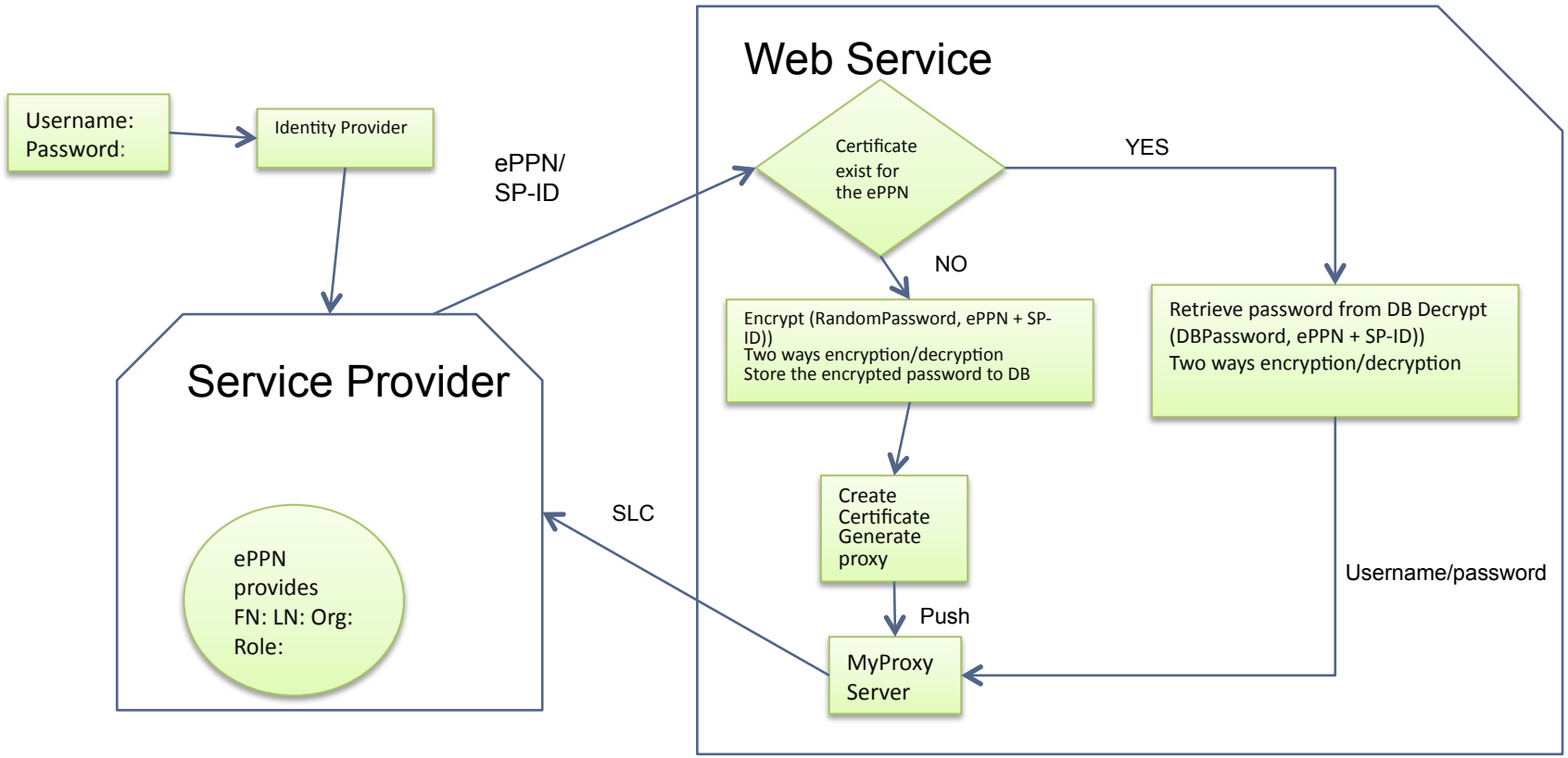
- Retrieve proxy from myproxy.ucgrid.org
- /u/local/apps/kepler/kepler.modules/build-area/kepler.sh
  -MyProxyFilePath /tmp/x509up_uid
  -GlobusTaskPath4Hoffman2
  /u/scratch/ppk/rosetta/rosetta-grid-workflow-hoffman2
  -GlobusTaskPath4Hydro
  /u/scratch/ppk/rosetta/rosetta-grid-workflow-hydro
  -ScaffoldPath4Hoffman2
  /u/home2/ppk/rosettatest/scaffold-hoffman2-one-scaffold
  -ScaffoldPath4Hydro
  /u/home2/ppk/rosettatest/scaffold-hydro-one-scaffold
  /u/local/apps/kepler/rosetta-workflow/Rosetta-Match-BChain-Design-Globus-TopLevel-Parallel-v1-18.xml

- TaskPaths are scratch directories for the run and ScaffoldPath is the location of input files for each cluster pre-divided by the user. Workflow directives are in the xml file

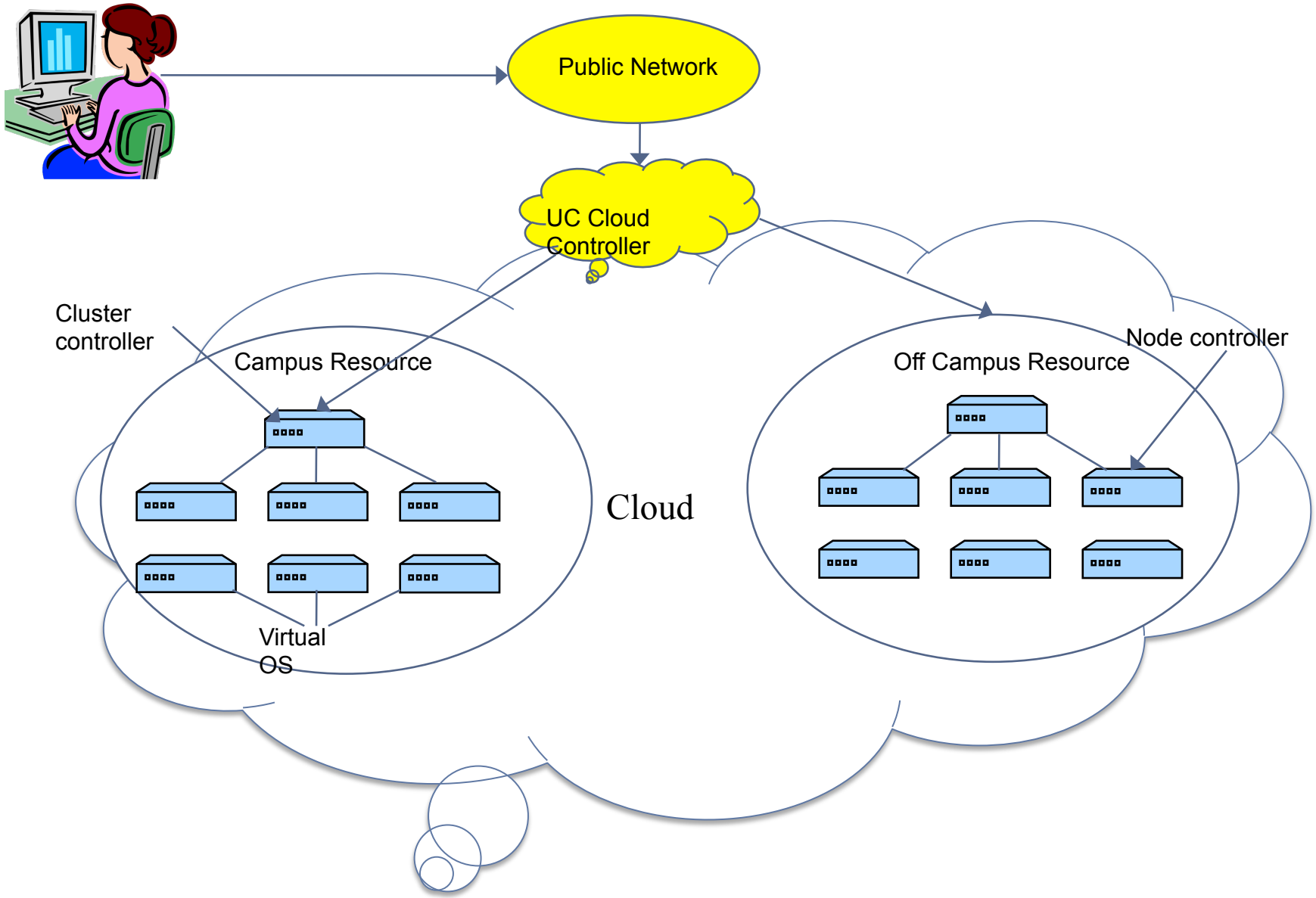
Experimental data on Hoffman2 cluster

Workflow name	Workflow Structure	Inputs	Main configurations	Job number	Execution time
Rosetta-MatchSGE-v2-4.xml	MatchSGEJob	226 Scaffold	1) Match command number for each job: 5 Generate 108,012 pdb files.	46 Match SGE jobs.	5.36 hours
Rosetta-MatchSGE-BChain-v1-6.xml	MatchSGEJob + BChain	226 Scaffold	1) Match command number for each job: 5 Generate 216,024 pdb files.	46 Match SGE jobs.	5.19 hours
Rosetta-DesignSGE-v1-8.xml	DesignSGEJob (each design includes cst, des and min)	10 Scaffold	1) Structure Number for Design Calculation: 10. 2) Command number for each job: 5 (cst) + 5 (des) + 50 (min) = 60. Generate 14691 pdb files.	286 Design SGE jobs	10.73 hours
Rosetta-MatchSGE-BChain-DesignSGE-v1-9.xml	SGEMatch, BChain, SGEDesign	10 scaffold	1) Command number for each Match job: 1. 2) Structure Number for Design Calculation: 10. 3) Command number for each Design job: 5 (cst) + 5 (des) + 50 (min) = 60.	10 Match Jobs and 286 Design Jobs.	6.70-17.01 hours.
Rosetta-DesignSGE-v1-8.xml	DesignSGEJob (each design includes cst, des and min)	226 Scaffold	1) Structure Number for Design Calculation: 100. 2) Command number for each job: 10 (cst) + 10 (des) + 1000 (min) = 1100. The whole cmd number could be 50 million and the whole execution could generate 7 million files: * 14691 * 10 = = 7,793,734,24. (2036 is the pdb file number for 10 Scaffold of Rosetta-MatchSGE-v2-4.xml)	Over 11,000 Design SGE jobs.	One job may last for over 18 hours. Estimated whole execution time : (108012/2036) * 10.73 * 10 = 7.79 months.

Workflow to Use Shibboleth Authentication to Sign X509 Certificate



UC Cloud Architecture



UC Grid Portal Contact Information

http://www.ucgrid.org

http://inca.ucgrid.org

atshpc@ats.ucla.edu

- Kepler: Jianwu Wang, SDSC, jianwu@sdsc.edu
- INCA, SDSC: Shava Smallen, ssmallen@sdsc.edu
- UGP: Kejian Jin kjin@ats.ucla.edu
- Rosetta: Seonah@chem.ucla.edu
- Proteomic: psouda@chem.ucla.edu
- UCLA: Prakashan Korambath, ppk@ats.ucla.edu
- UCSB: Paul Weakliem, weakliem@cnsi.ucsb.edu
- UCR: Bill Strossman, wlliam.strossman@ucr.edu
- UCI: Harry Mangalam, harry.mangalam@uci.edu