Graduate School of Biomedical Sciences, Dept. of Biochemistry
UT Health Science Center™
San Antonio

Borries Demeler, Ph.D.

Biophysics and HPC:

The UltraScan XSEDE Science Gateway

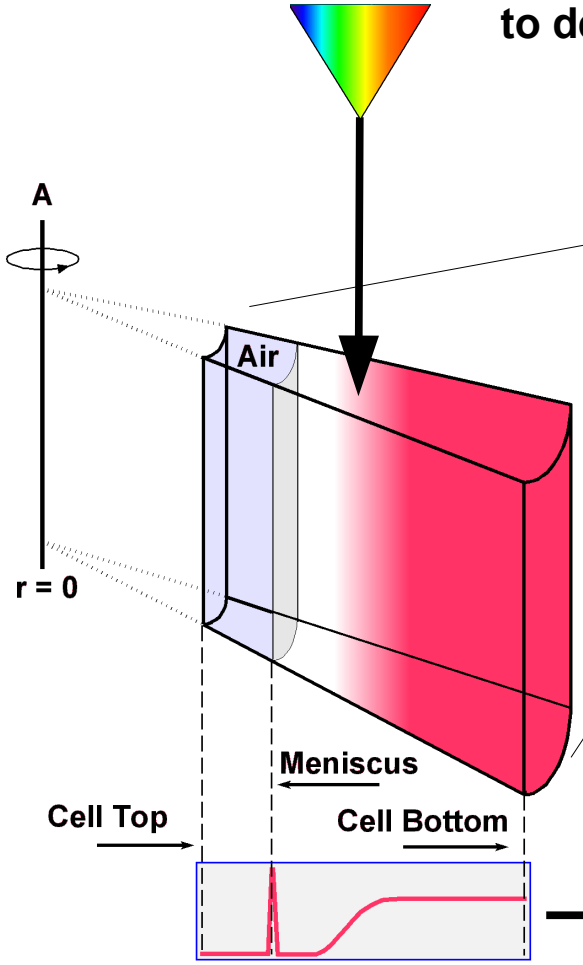
A Science Gateway for Biophysical Analysis

Goals:

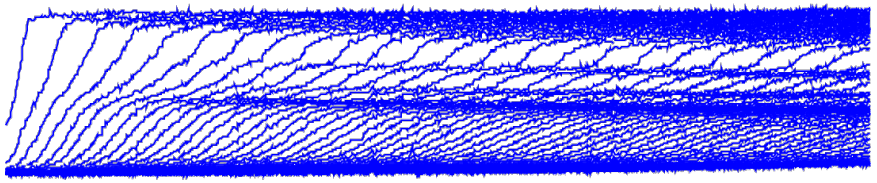
- *Support analysis of experiments on molecules in solution environments*
  - *Analytical ultracentrifugation (AUC)*
  - *Laser Light Scattering (LS)*
  - *Small angle X-ray/neutron scattering (SAXS/SANS)*
- *Provide highest possible resolution in the analysis – requires HPC*
- *Offer a flexible model for multiple optimization methods*
- *Integrate a variety of HPC environments into a uniform user experience*
- *Must be easy to learn and use – users should not have to be HPC experts*
- *Support a large number of users and analysis instances simultaneously*
- *Support data sharing and collaborations*
- *Easy installation, easy maintenance*
- *Robust and secure multi-user/multi-role framework*
- *Provide check-pointing and easy to understand error messages*
- *Support for multiple grid middleware (GRAM4, GRAM5, Unicore...)*
- *Fast turnaround to support serial workflows (model refinement)*

A Science Gateway for Biophysical Analysis

Challenge: Model analytical ultracentrifugation experiments to describe molecules under dynamic solution conditions: Computationally expensive!



Multiple recordings over time...



A Science Gateway for Biophysical Analysis

$$L(s,D): \quad \left(\frac{\partial C}{\partial t}\right)_r = \frac{-1}{r}\frac{\partial}{\partial r}\left[s\,\omega^2 r^2 C - D r \frac{\partial C}{\partial r}\right]_t$$

$$MIN: \sum_{i=1,j=1}^{r,t}\left[Y_{i,j} - \sum_{k=1}^{n}\left(L_k(s_k, D_k)_{i,j}\right)\right]^2$$

The experiment is described by the Lamm equation, which is solved with FEM

Cao W and Demeler B. Modeling analytical ultracentrifugation experiments with an adaptive space-time finite element solution of the Lamm equation. (2005) Biophys J. 89(3):1589-602.

Cao, W and Demeler B. Modeling Analytical Ultracentrifugation Experiments with an Adaptive Space-Time Finite Element Solution for Multi-Component Reacting Systems. Biophys. J. (2008) 95(1):54-65

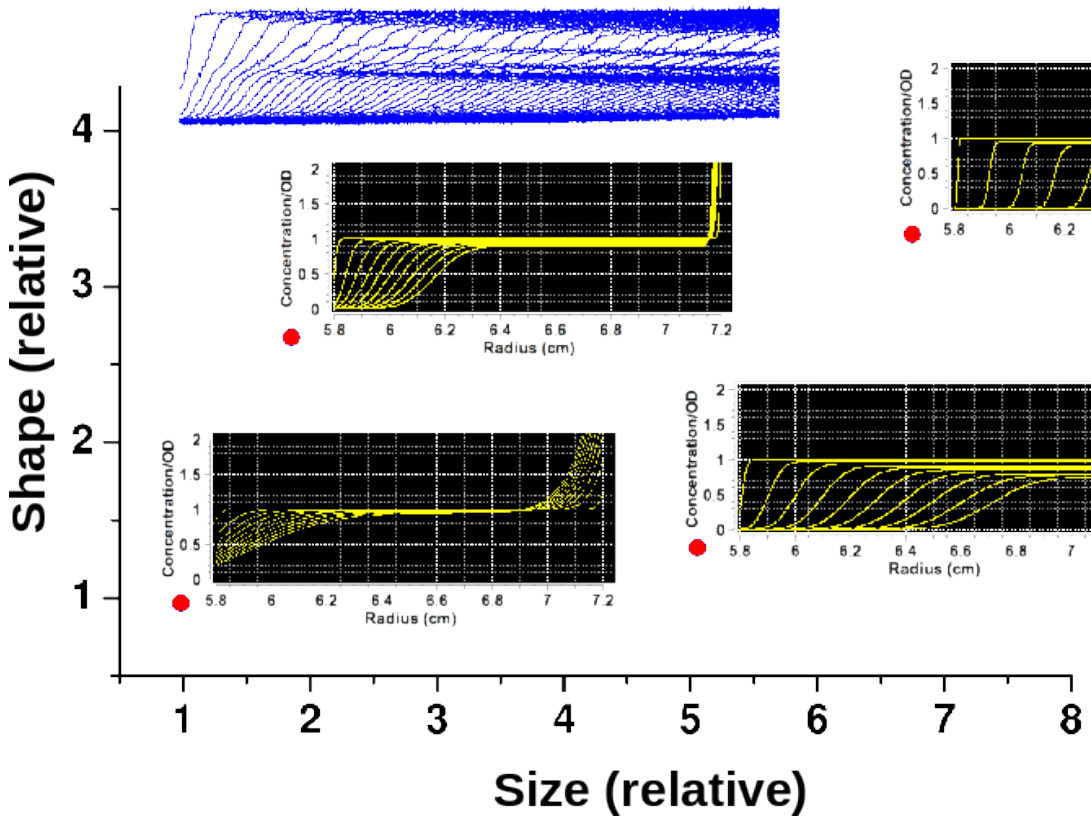
The optimization problem is solved with parallelized 2DSA, GA, MC

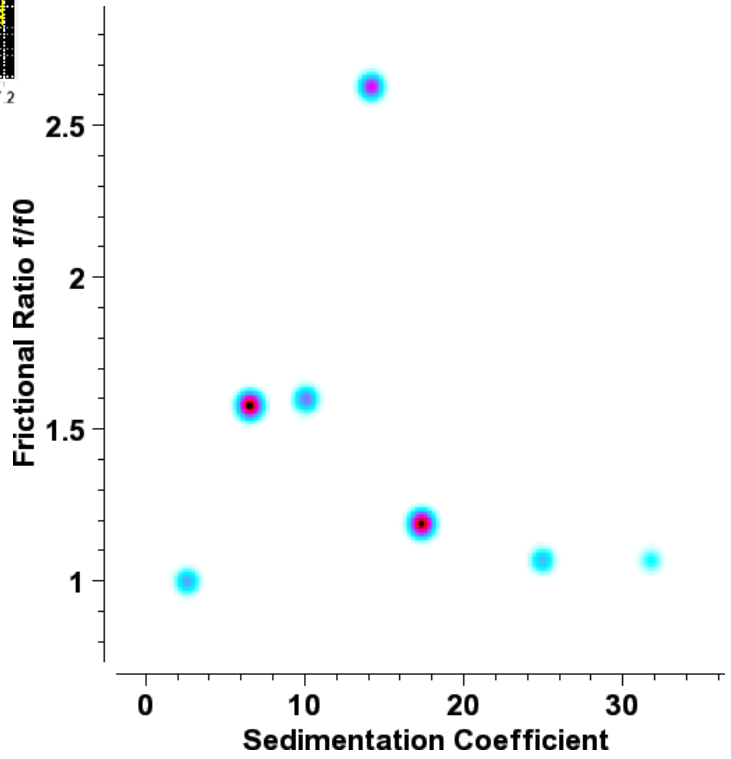
Brookes E, Cao W, Demeler B A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. Eur Biophys J. (2010) 39(3):405-14.

Brookes E, Demeler B. Parsimonious Regularization using Genetic Algorithms Applied to the Analysis of Analytical Ultracentrifugation Experiments. GECCO Proceedings ACM 978159593-697-4/07/0007 (2007)

Demeler B and E. Brookes. Monte Carlo analysis of sedimentation experiments. Colloid Polym Sci (2008) 286(2) 129-137

The analysis provides information about the number of components in solution, their relative concentrations, their sedimentation and diffusion properties, molar mass, anisotropy (rel. shape), interaction coefficients, and reaction rate constants

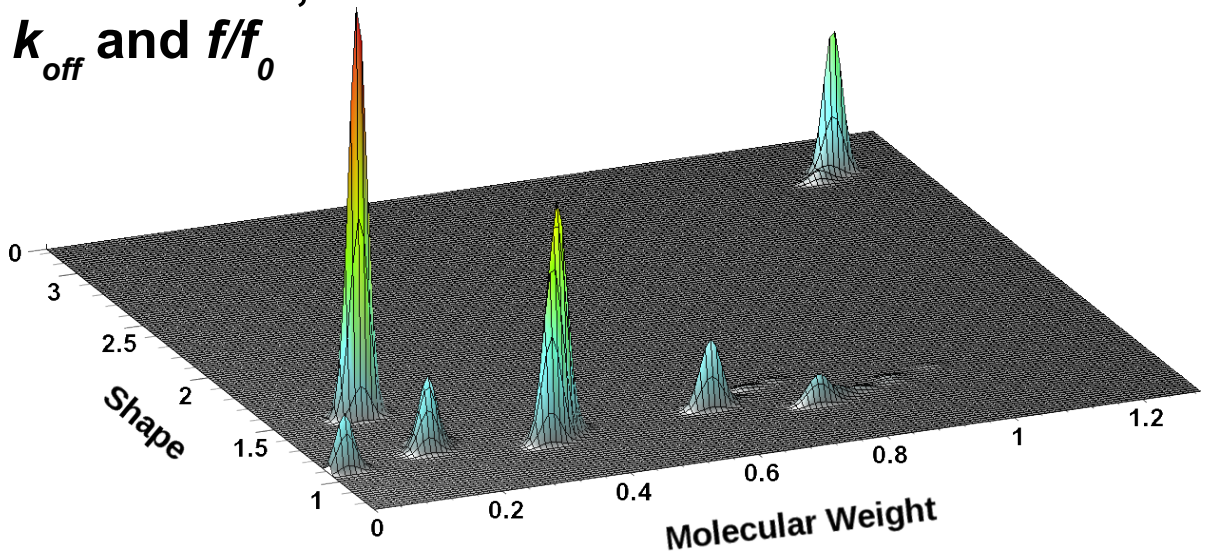




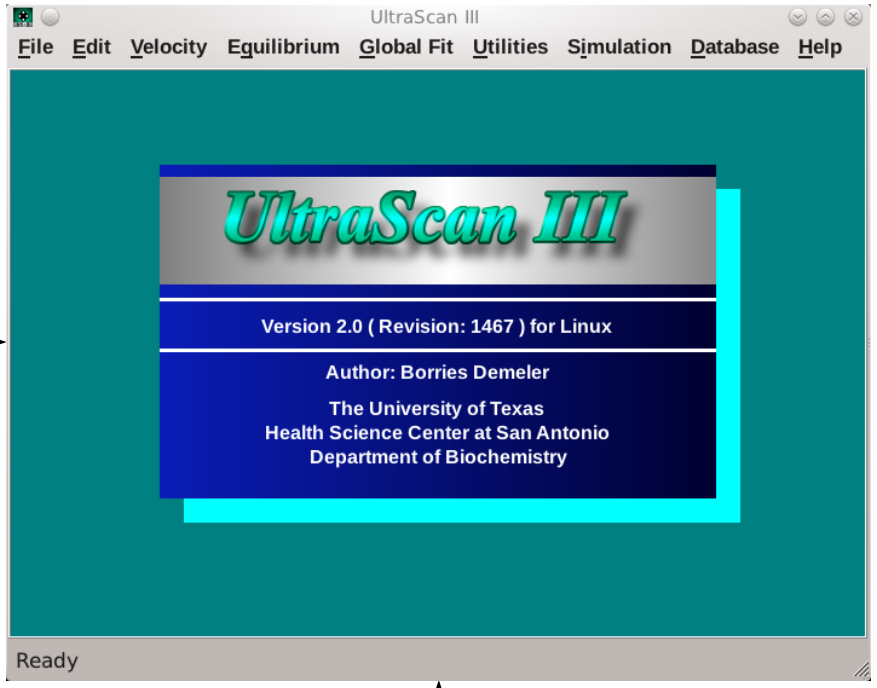
Find: partial concentration, *s, D, MW, Kd,* k_{off} and f/f_0

Optimization:

2D-Spectrum Analysis
Genetic Algorithms
Monte Carlo Analysis



UltraScan Organization (open source, multi-platform)



User

Web Server

MySQL DB

US LIMS

UTHSCSA Jacinto
Terascale storage

Airavata

Globus Network
UNICORE

XSEDE, etc

High Performance
Computing Clusters

UltraScan (US) components and 3rd party software:

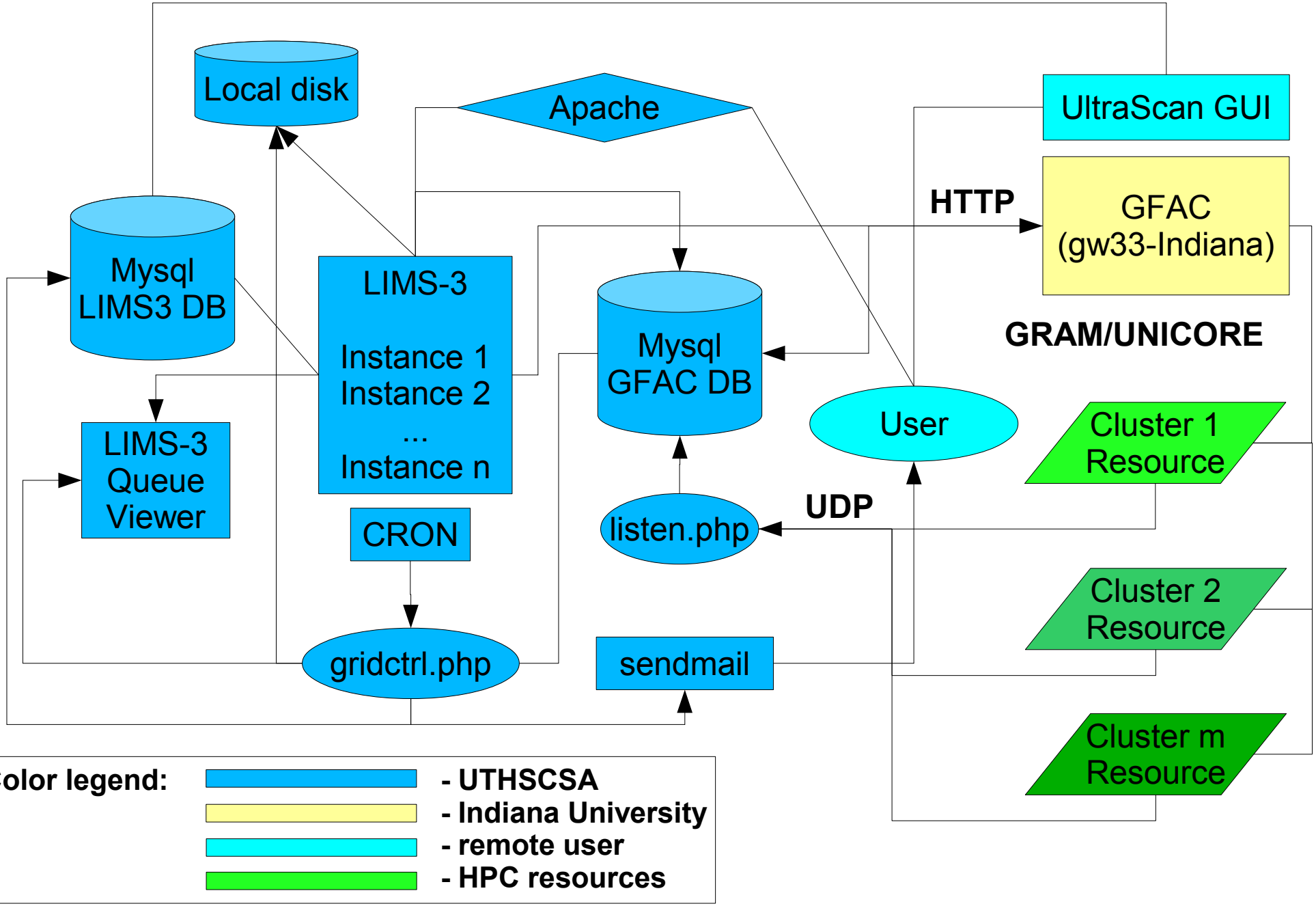
1. US-GUI: (Linux, Mac, Windows): used for data processing, editing, uploading, visualization, report generation (C++/Qt).
2. US-LIMS: web-based interface to the MySQL database. Gives the user web-based access to their data, visualizations and analysis reports (PHP).
3. US-HPC: includes finite element solver for multiple models and performs all optimization algorithms (2DSA, GA, Monte Carlo)
4. US-Gridcontrol: PHP daemon to manage job submission, communications and reporting from US-HPC,
5. MySQL: houses all LIMS data, a separate GFAC DB holds job status info. Each institution has a separate database and separate US-LIMS instance. All communications with MySQL are performed through stored procedure calls.
6. GFAC/Airavata: accepts job requests and brokers them over the appropriate grid middleware to the requested supercomputer, controls job status in GFAC DB.
7. GRAM/UNICORE grid middleware
8. SSL communications: All communications and data transfers are SSL encrypted. All database access and LIMS access requires authentication.
9. Statistic scripts: queries database and reports usage and provides details for XSEDE reporting.

Data Flow:

1. User acquires data and stores it on a local computer.
2. After authentication in US-GUI, user uploads data (converted to efficient binary format) together with all meta data (XML) information to the LIMS SQL database.
3. Data editing: edited data is used by all analyses in US for further processing.
4. User logs into US-LIMS website, selects data, analysis method, analysis parameters, and HPC resource, then submits the analysis request. LIMS places the data and request into a special GFAC database, and notifies GFAC.
5. GFAC component brokers request for desired HPC resource, grabs data from GFAC database, and places the job into the job queue.
6. US-HPC processes analysis, sending UDP status updates which are processed by US-Gridcontrol
7. User monitors job queue which is periodically updated by US-Gridcontrol. When job is finished, user is also sent an e-mail, and the result (a universal model in XML) is deposited in the MySQL DB associated with the user's US-LIMS instance.
8. User visualizes the results with the US-GUI and prepares items to be stored in the report. Results and reports are placed into MySQL DB.
9. User can proceed with several iterations of result refinement (GA, MC)
10. User can share the data with a collaborator who has an account on the same LIMS. User or collaborator can log into US-LIMS and review the results.

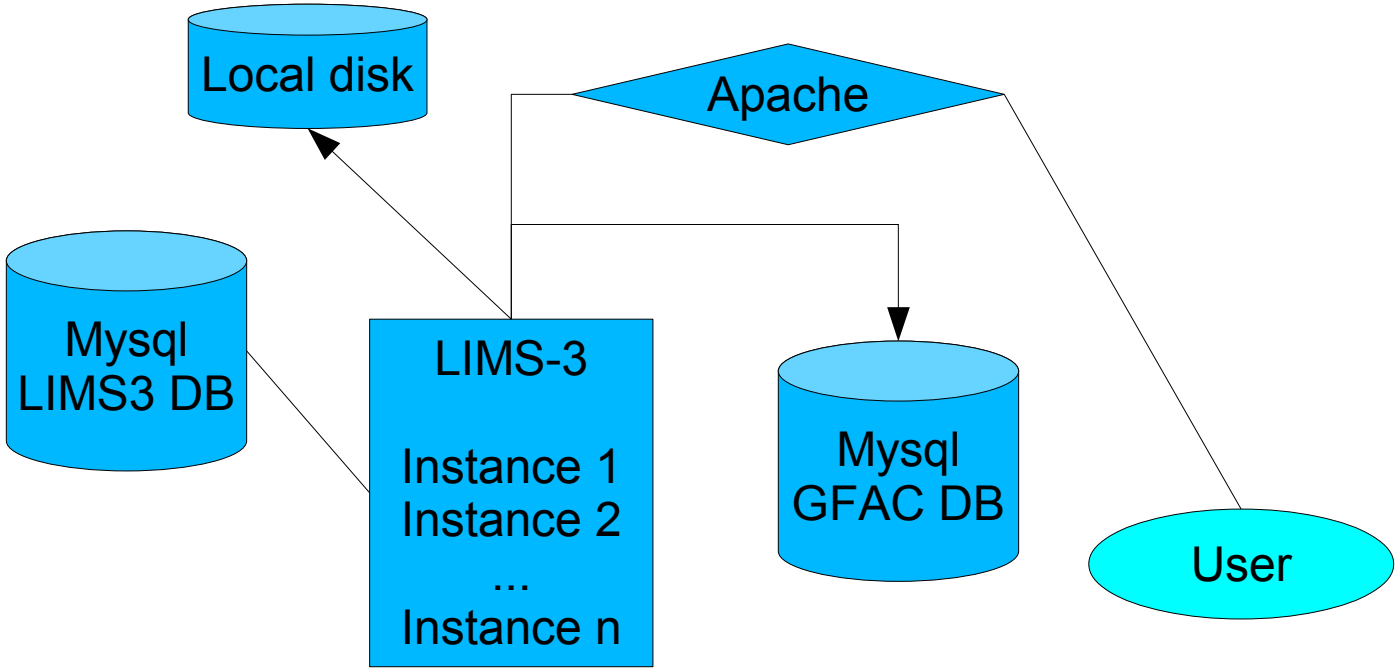
UltraScan LIMS3/GFAC Workflow Overview

(GFAC = Generic Factory Component of Apache Airavata)



UltraScan LIMS3/GFAC Workflow Overview

(Step 1: User logs into LIMS and requests their data, sets analysis parameters, job is retrieved from DB and packaged for submission, job request gets stored in GFAC DB)

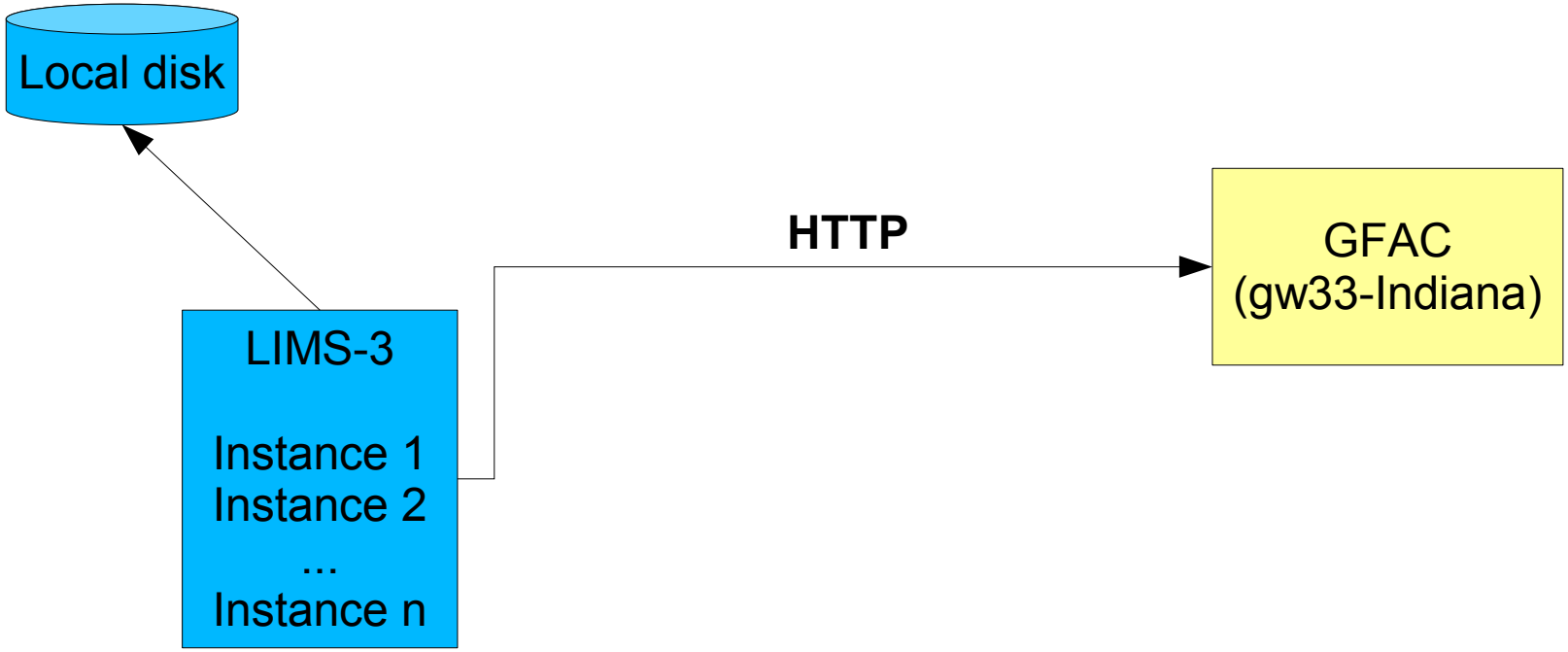


Color legend:

- **UTHSCSA**
- **Indiana University**
- **remote user**
- **HPC resources**

UltraScan LIMS3/GFAC Workflow Overview

(Step 2: LIMS packages data and job parameter on the local storage and sends notification of pending job to GFAC)

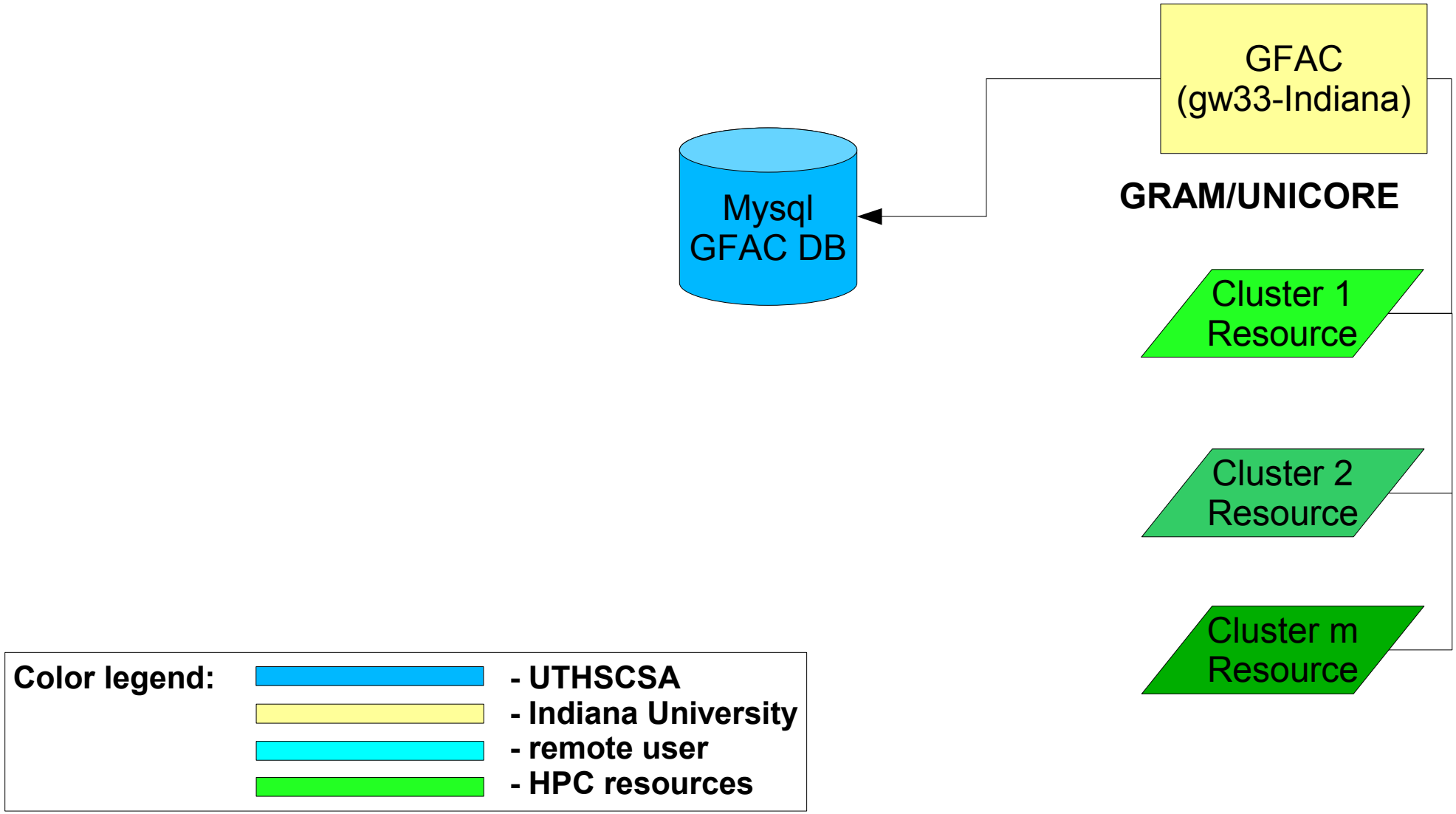


Color legend:

- **UTHSCSA**
- **Indiana University**
- **remote user**
- **HPC resources**

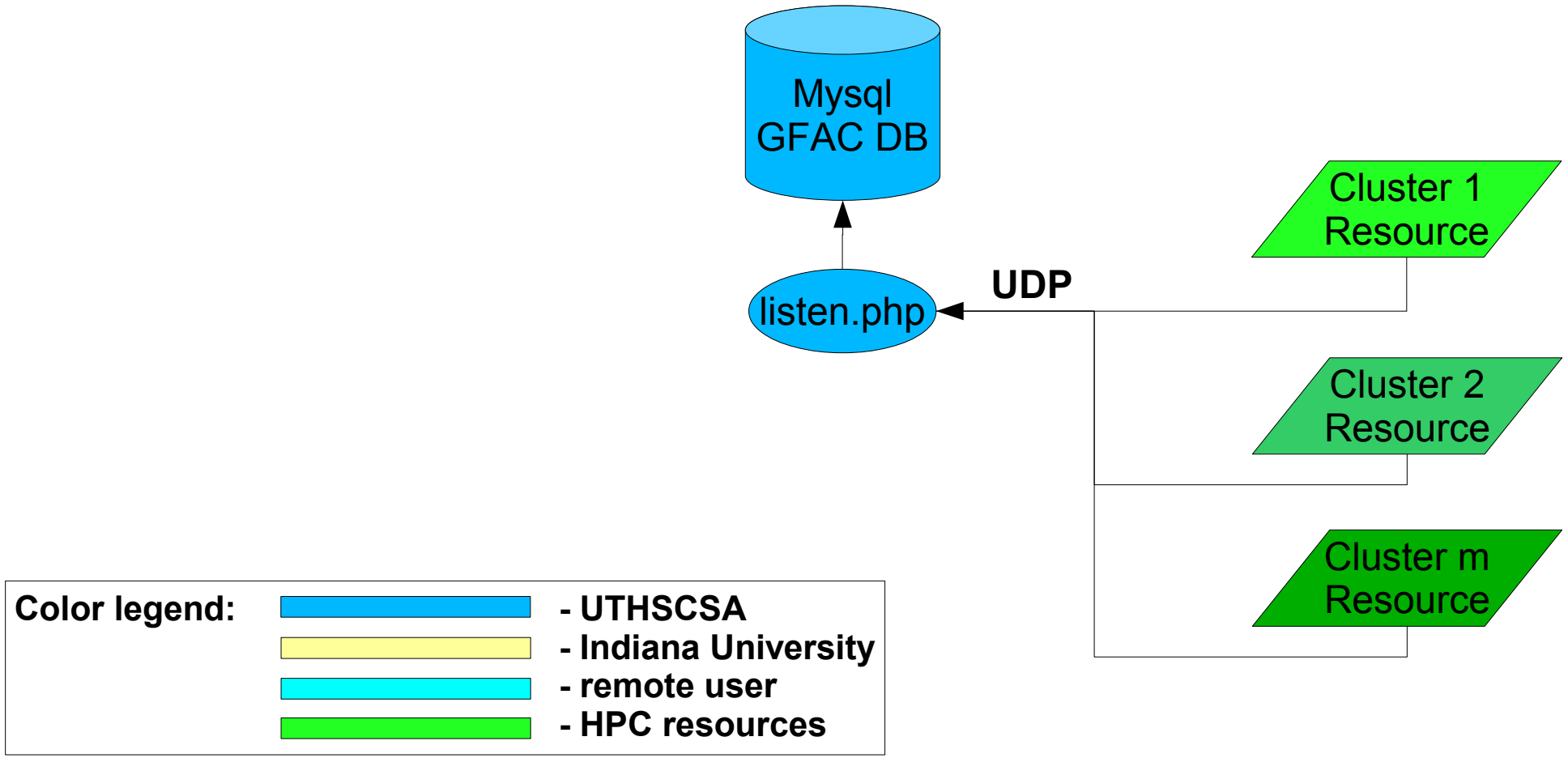
UltraScan LIMS3/GFAC Workflow Overview

(Step 3: GFAC grabs data from GFAC DB, and queues the job on the requested resource)



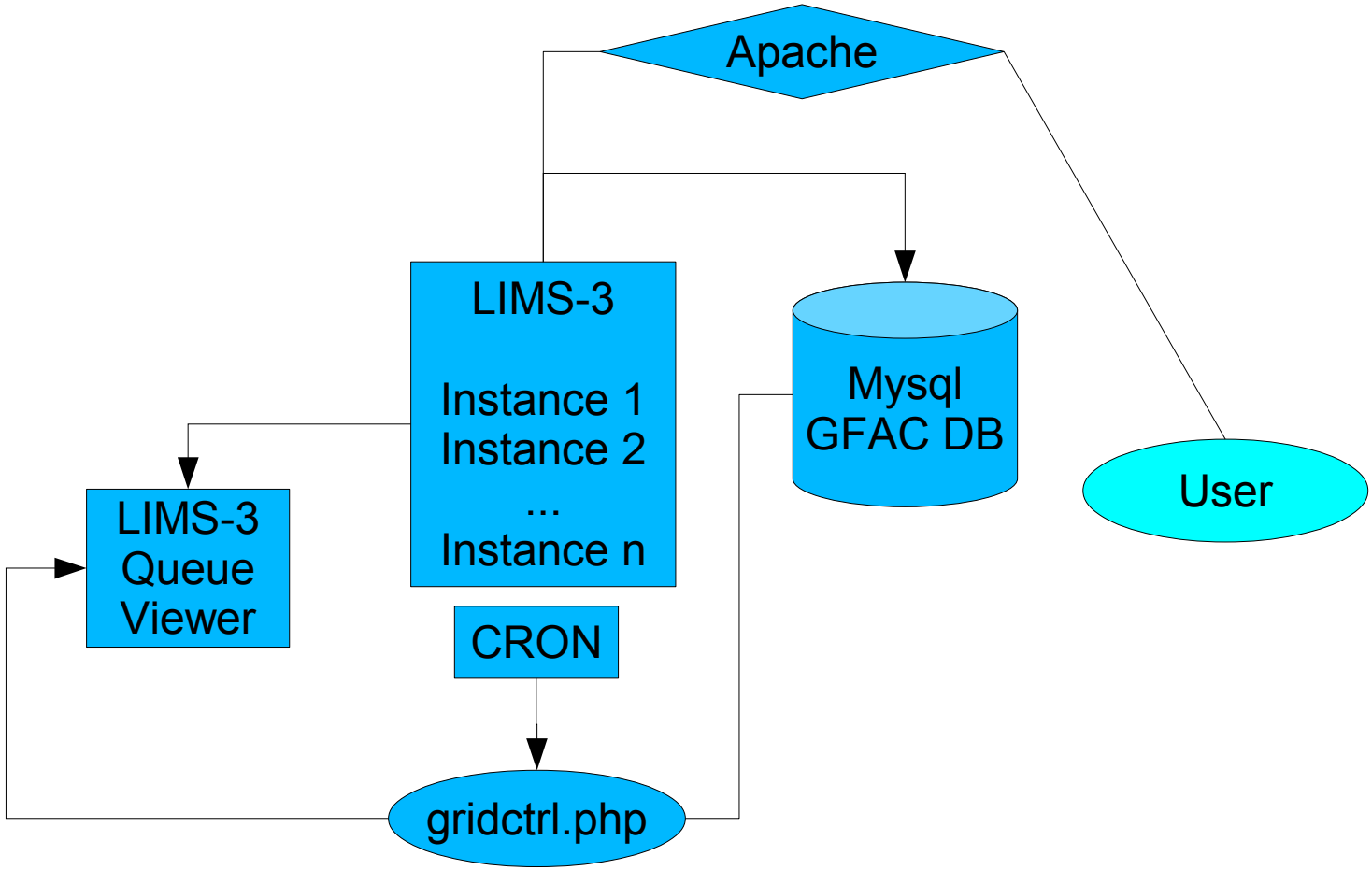
UltraScan LIMS3/GFAC Workflow Overview

(Step 4: US-HPC sends job status over UDP to a daemon which updates GFAC DB)



UltraScan LIMS3/GFAC Workflow Overview

(Step 5: cron triggers gridcontrol script periodically to check each job status in the GFAC DB, updates the Queue Viewer. User checks queue viewer through LIMS)

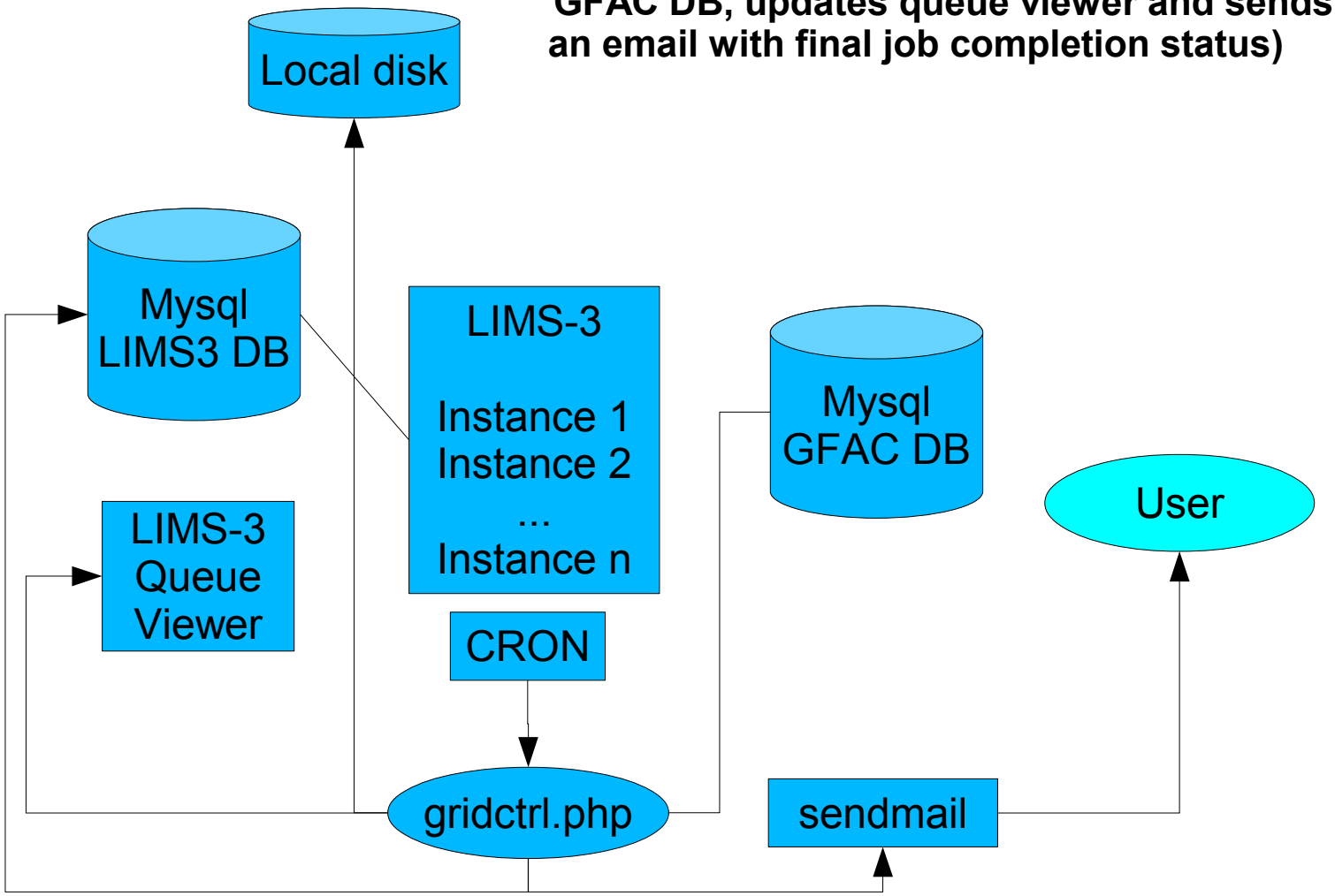


Color legend:

- **UTHSCSA**
- **Indiana University**
- **remote user**
- **HPC resources**

UltraScan LIMS3/GFAC Workflow Overview

(Step 6: cron triggers gridcontrol to check job list on GFAC DB for completed jobs, moves results to LIMS DB, copies stderr/out results to local disk, deletes completed jobs from GFAC DB, updates queue viewer and sends user an email with final job completion status)

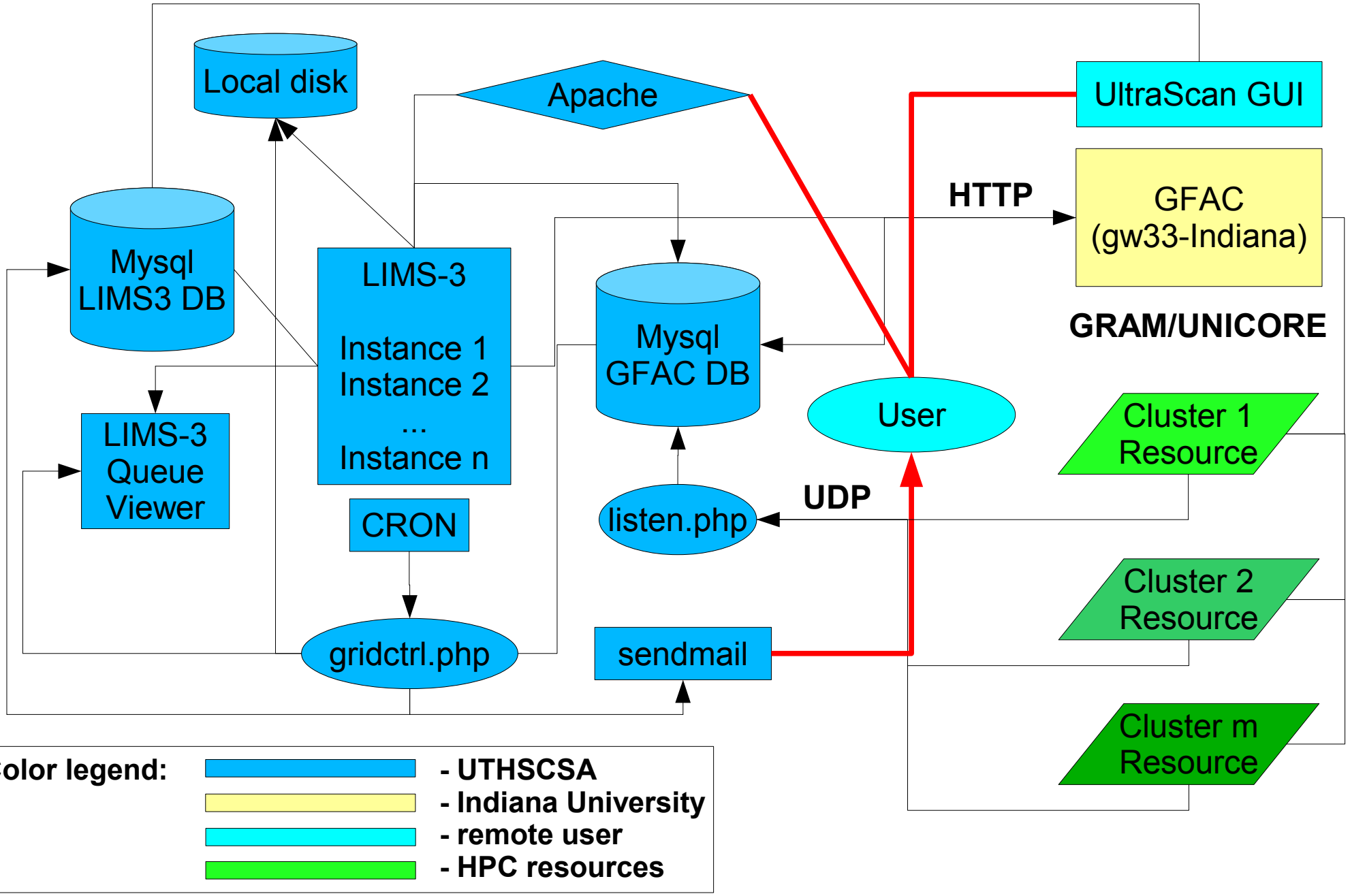


Color legend:

- **UTHSCSA**
- **Indiana University**
- **remote user**
- **HPC resources**

UltraScan LIMS3/GFAC Workflow Overview

(User only interacts with web service or uses US-GUI to visualize final results)



A Science Gateway for Biophysical Analysis

Current usage statistics (in SUs) for the period of January-November 2012:

Alamo (UTHSCSA)	BCF (UTHSCSA)	Ranger (TACC)	Lonestar (TACC)	Trestles (SDSC)	Total SUs	Number of Investigators
438,471	14,951	696,807	337,902	4,487	1,492,618	98

- There are over 40 institutions in 10 countries actively using UltraScan/LIMS
- Implemented currently on 5 HPC platforms, including one commercial installation (non-public)

Ongoing Projects:

- Integration of SAXS/SANS modeling
- Integration of Light Scattering modules
- Integration of Molecular Dynamics (DMD)
- Integration of Jülich Supercomputing Center (Germany)
- Integration of Multi-wavelength optics (500-1000 fold higher data density)

In the planning stage:

- Integration of CFA data acquisition

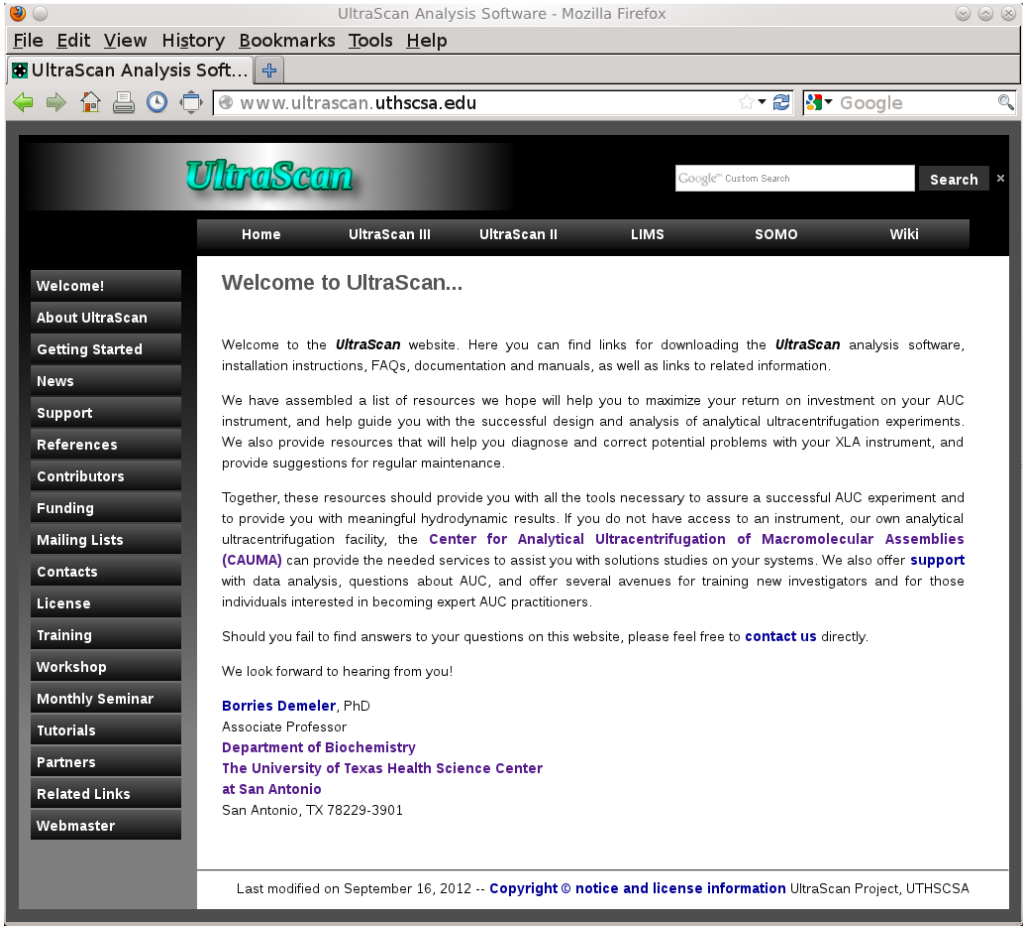
Links:

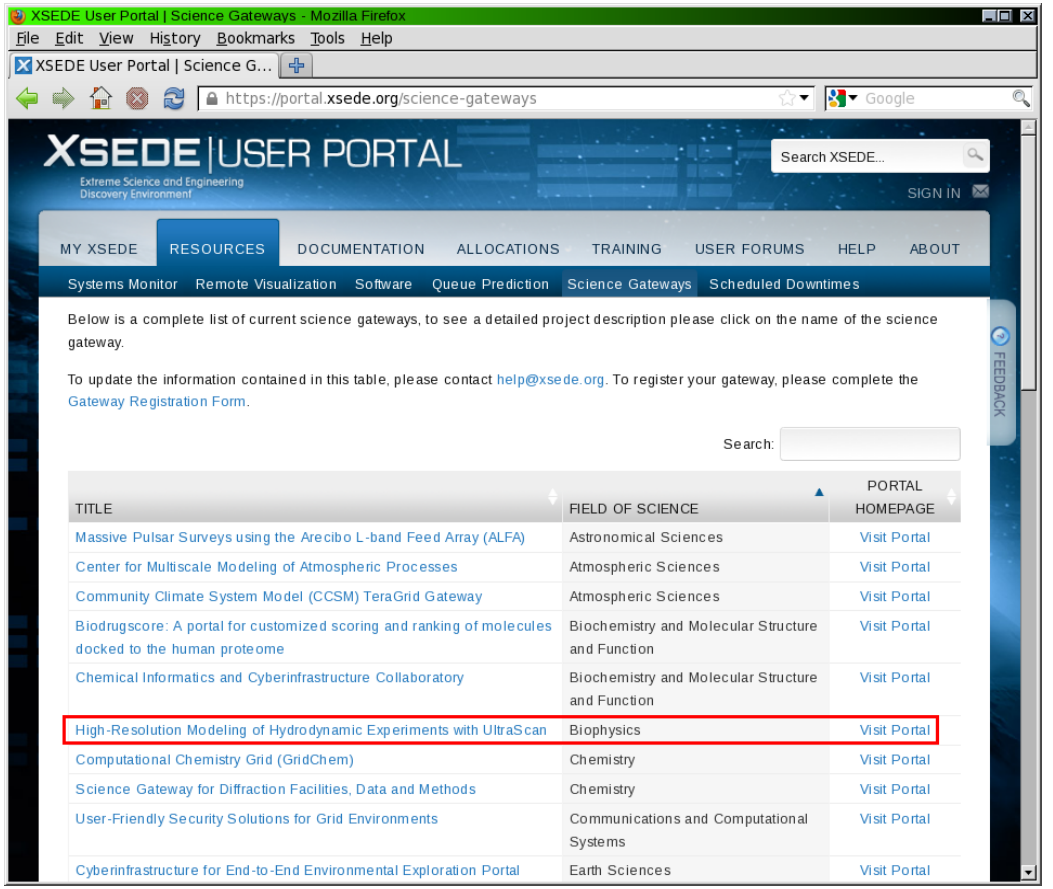
Ultrascan:

http://www.ultrascan.uthscsa.edu

XSEDE Science Gateway:

https://portal.xsede.org/science-gateways





A copy of this presentation:

http://www.demeler.uthscsa.edu/SciSoft2012/Presentation-121712.pdf

Acknowledgements:

UTSA

Dept. of
Applied Mathematics

Weiming Cao

Indiana University

Marlon Pierce
Suresh Marru
Raminder Singh

WE MAKE LIVES BETTER
UT HEALTH SCIENCE CENTER™
SAN ANTONIO

Demeler Lab:

Emre Brookes
Virgil Schirf
Jeremy Mann
Gary Gorbet
Dan Zollars
Bruce Dubbs

Funding:

NIH/NCRR
SACI (NCI)
National Science Foundation
NSF Teragrid (XSEDE)
San Antonio Life Science Institute
Howard Hughes Medical Institute
UT Permanent University Fund