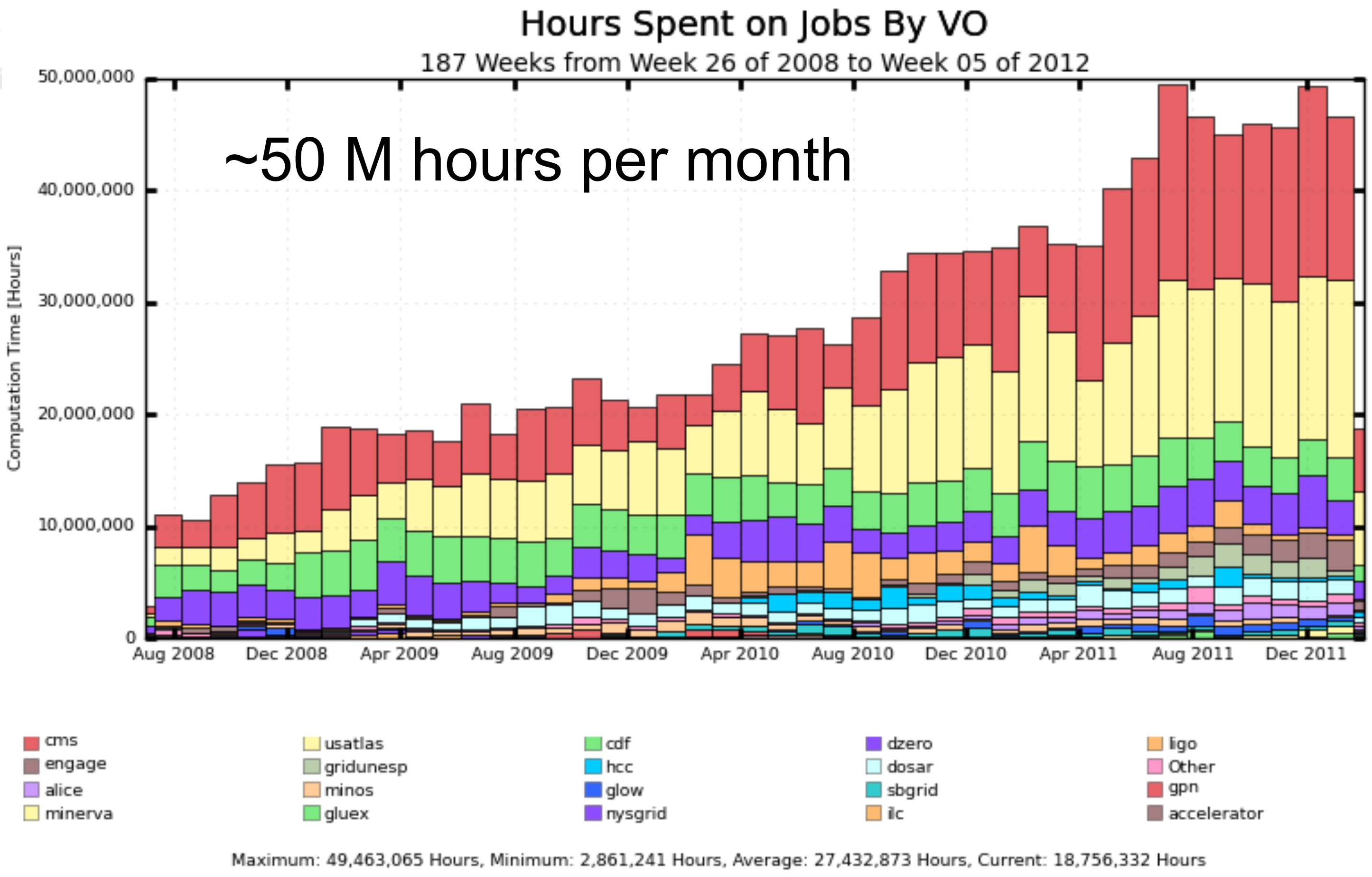
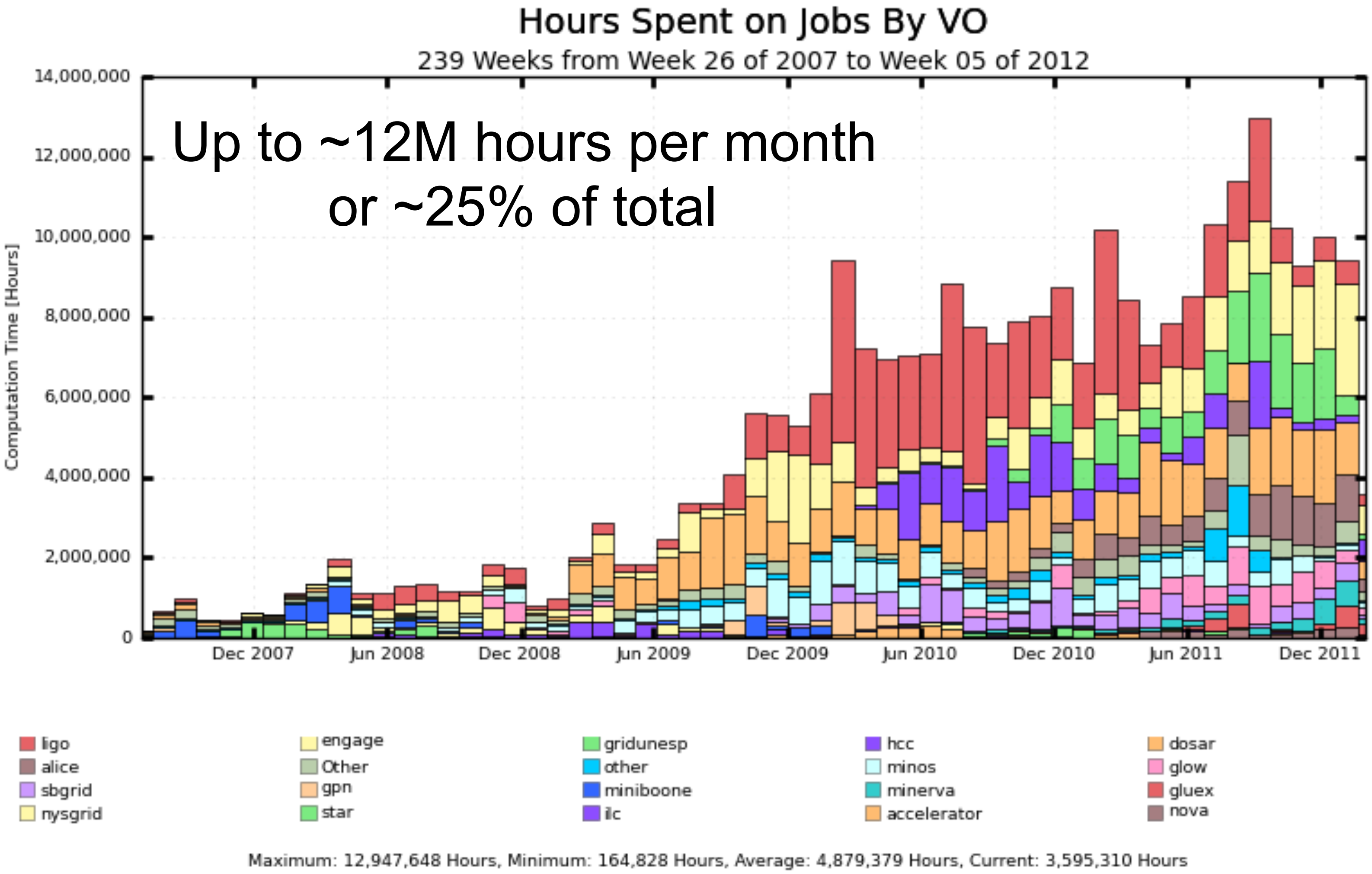
Science on OSG



~50 M hours per month

OSG All Hands Meeting 2012
Frank Würthwein (UCSD)

Let’s exclude HEP and pick topics from the rest



Hours Spent on Jobs By VO

239 Weeks from Week 26 of 2007 to Week 05 of 2012

Up to ~12M hours per month or ~25% of total

Maximum: 12,947,648 Hours, Minimum: 164,828 Hours, Average: 4,879,379 Hours, Current: 3,595,310 Hours

Themes for this talk

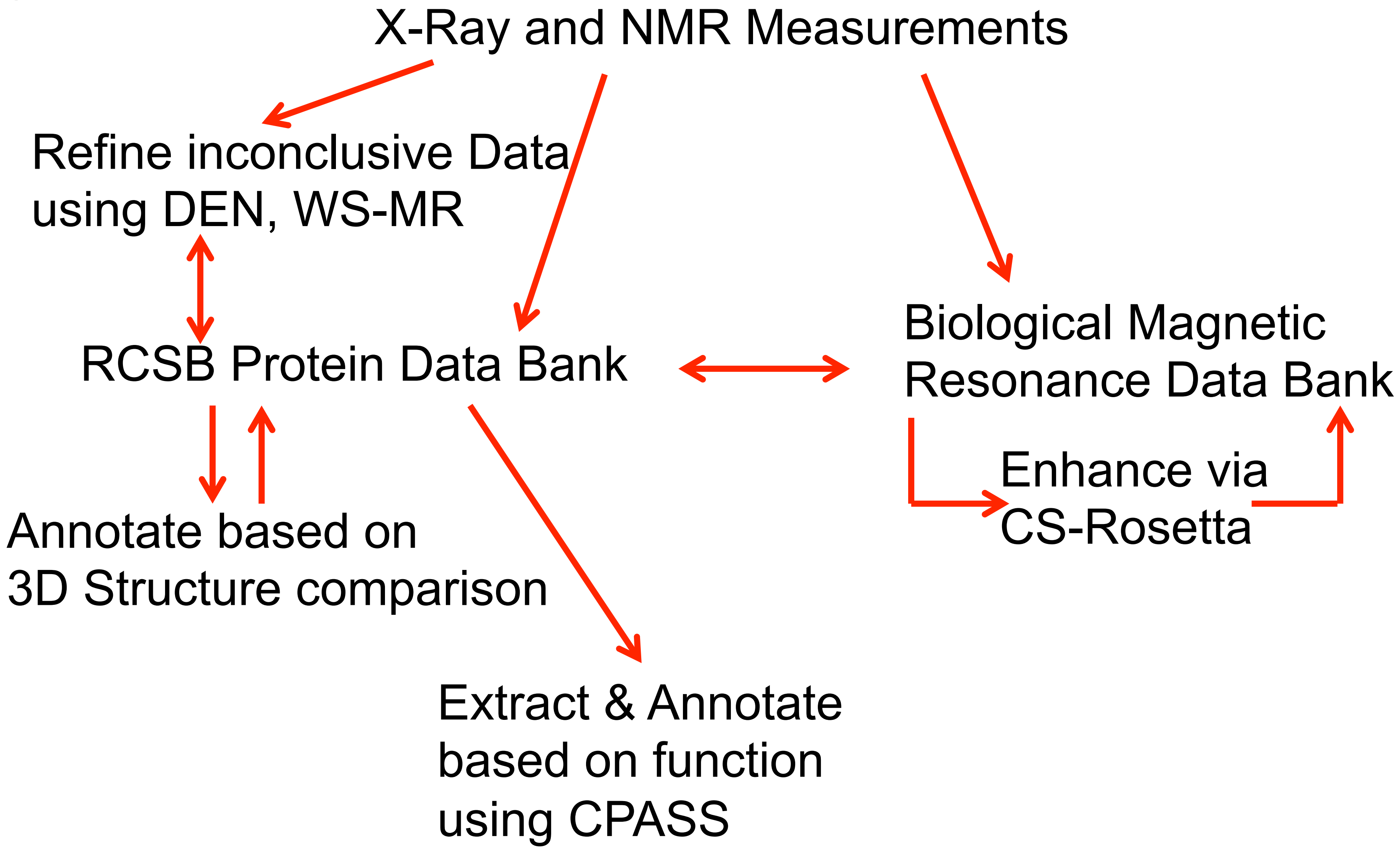
- One Science Theme
  - Protein Structure
- Community Theme
  - NEEShub, NanoHub, SBGrid, Engage, et al.
- Campus & Regional CI
  - GridUNESP, HCC, DOSAR, GLOW, GPN, NYSGrid, … and more …

Don’t have time to talk about all of these.
Will select a few examples to make my case …

Personal Agenda with this Talk

- OSG provides access to resources at O(100) institutions in the US.
- Many of you are directly involved in making this possible.
- **I want to motivate you to go out and find Scientists on your campus with Science that can benefit from Distributed High Throughput Computing (DHTC).**
- Many of the examples I have in this talk came about because people like you reached out to their "local" community.
- The prevailing paradigm is to submit locally (or via a portal), and then send the overflow to the OSG via glideinWMS.

Protein Structure From Experiment



X-Ray and NMR Measurements

Refine inconclusive Data using DEN, WS-MR

RCSB Protein Data Bank

Biological Magnetic Resonance Data Bank

Enhance via CS-Rosetta

Annotate based on 3D Structure comparison

Extract & Annotate based on function using CPASS

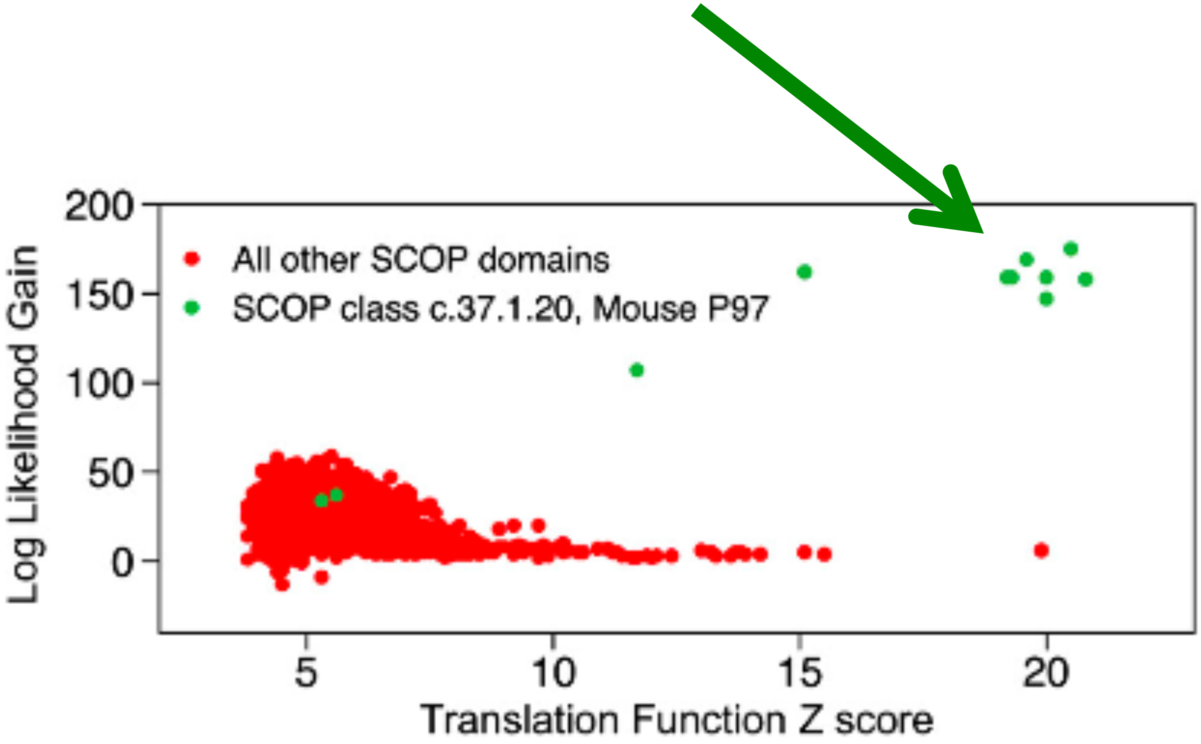
X-ray Crystallography:
Wide Search Molecular Replacement

WS-MR success- fully identifies the closest structural homologues from a large family of candidates.

WS-MR is used to determined new X-ray structures using phasing information from identified structural homologues.

Individual WS-MR search is performed on ~100,000 domains derived from the Protein Data Bank.

Identified "hits"



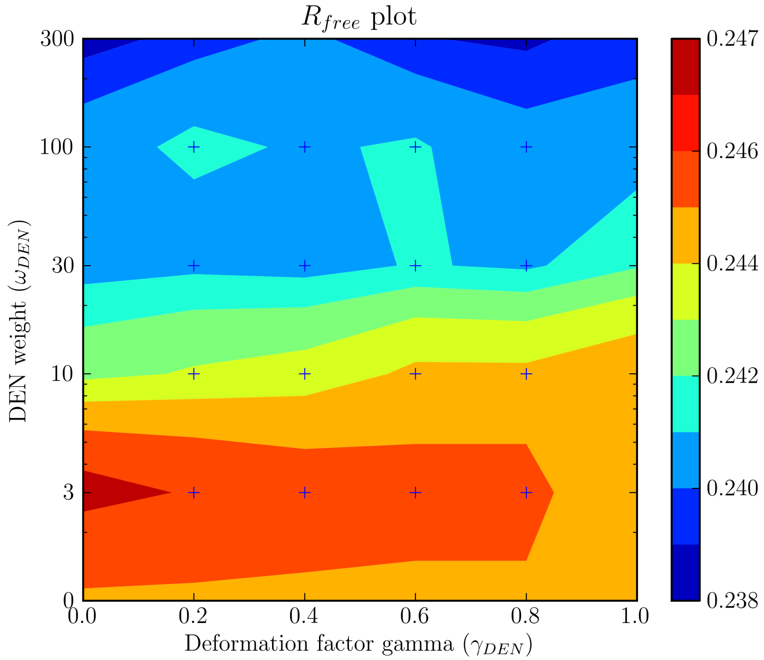
Stokes-Rees, I., and Sliz, P. (2010). Protein structure determination by exhaustive search of Protein Data Bank derived databases. PNAS. *107*, 21476–21481.

X-ray Crystallography:
Low Resolution Structure Refinement

Refinement of structures at resolutions lower than 3.5 A is known to be difficult and error-prone.

DEN refinement is used to refine low resolution structures using a high resolution reference model.

SBGrid provides DEN and WS-MR as a **webservice** via its portal and uses OSG CI via **glideinWMS** as backend. **HMS Orchestra** cluster is now also integrated with OSG and supports portal infrastructure.



An Rfree heat map of results from the Notch protein DEN optimization.

O'Donovan, D.J., Stokes-Rees, I., Nam, Y., Blacklow, S.C., Schröder, G.F., Brunger, A.T., and **Sliz, P.** (2012). A grid-enabled web service for low-resolution crystal structure refinement. Acta Cryst. D*68*, 261–267.

Choi, S.H., Wales, T.E., Nam, Y., **O'Donovan, D.J.**, **Sliz, P.**, Engen, J.R., and Blacklow, S.C. (2012). Conformational locking upon cooperative assembly of notch transcription complexes. Structure *20*, 340–349.

Calculating Pairwise Similarities of Proteins in RCSB Protein Data Bank (PDB)

- Detection of similarities in protein structure is important to infer functional and evolutionary relationships between protein families.
- A 3D structure comparison of 140 Million pairs is under way, resulting in annotation of PDB information that is then available to the biomedical science community worldwide.
- PDB is used by 200,000 unique scientists per month.

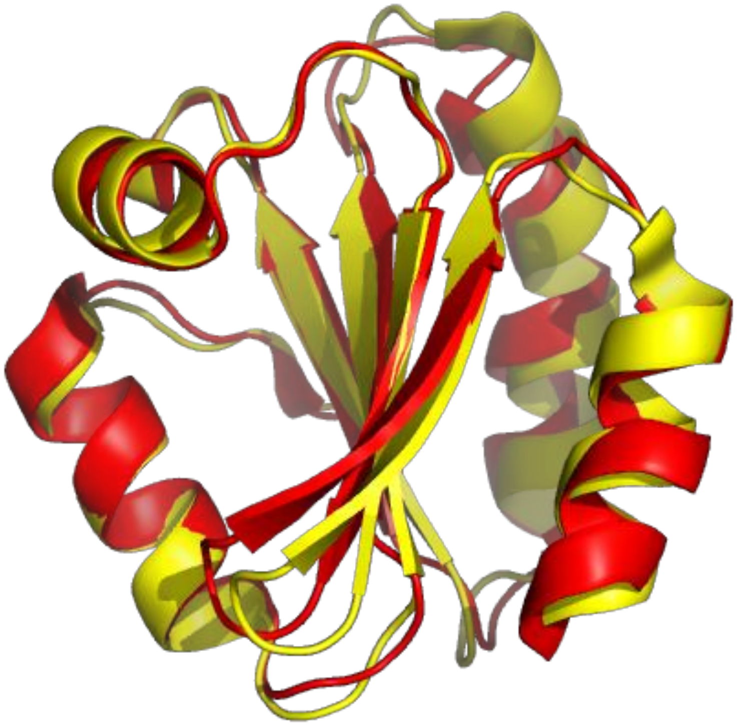


Figure 1: Structural alignment of thioredoxins from humans and the fly Drosophila melanogaster. The proteins are shown as ribbons, with the human protein in red, and the fly protein in yellow. Generated from PDB IDs 3TRX and 1XWC. (Image taken from http://en.wikipedia.org/wiki/Structural_alignment)

A.Prlic (PDB/SDSC), C.Bizon (RENCI)

Protein structure determination using CS-Rosetta at the BioMagResBank

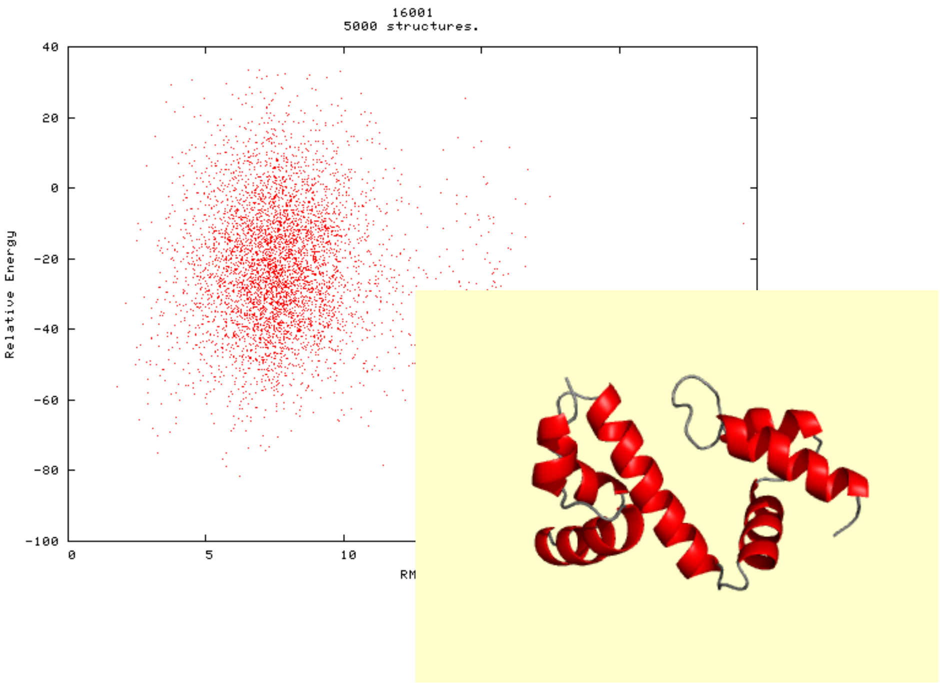
CS-Rosetta uses Monte Carlo simulations based on Nuclear Magnetic Resonance (NMR) data and a database of known structures to predict the structure of proteins.

Determining these structures through lab work alone is very time consuming and not always possible.

The question: how accurate is CS-Rosetta and which types of proteins does it work best on?

OSG allows us to run CS-Rosetta simulations on the entire BMRB database (a public database which holds NMR data). This allows us to study the predictive validity of CS-Rosetta as well as provide additional information for users studying proteins in our database.

Without OSG, there would be more new data submitted to the BMRB than CPU hours available to run the simulations and we would never be able to complete this project!

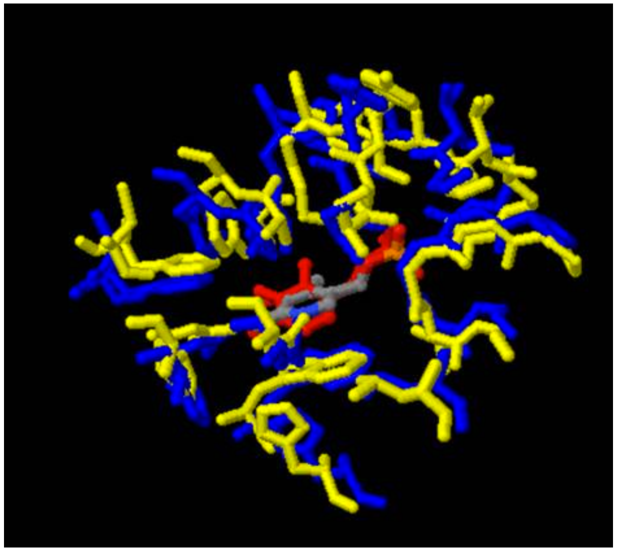


Jon Wedell - BioMagResBank

CPASS

Comparison of Protein Active-Site Structures (CPASS) annotates Protein data base info for proteins of unknown function with proteins of known function that share similar ligand-binding sites.

Underlying premise: ligand binding sites are more evolutionary stable than the underlying protein. Function can thus be inferred despite evolution of protein structure.



CI need:
Comparison against protein info in PDB.
CI Strategy:
Submit to local cluster, and elevate into OSG via glideinWMS as needed.

"Overflow into OSG"

R.Powers, A.Caprez et al. (see poster for details)

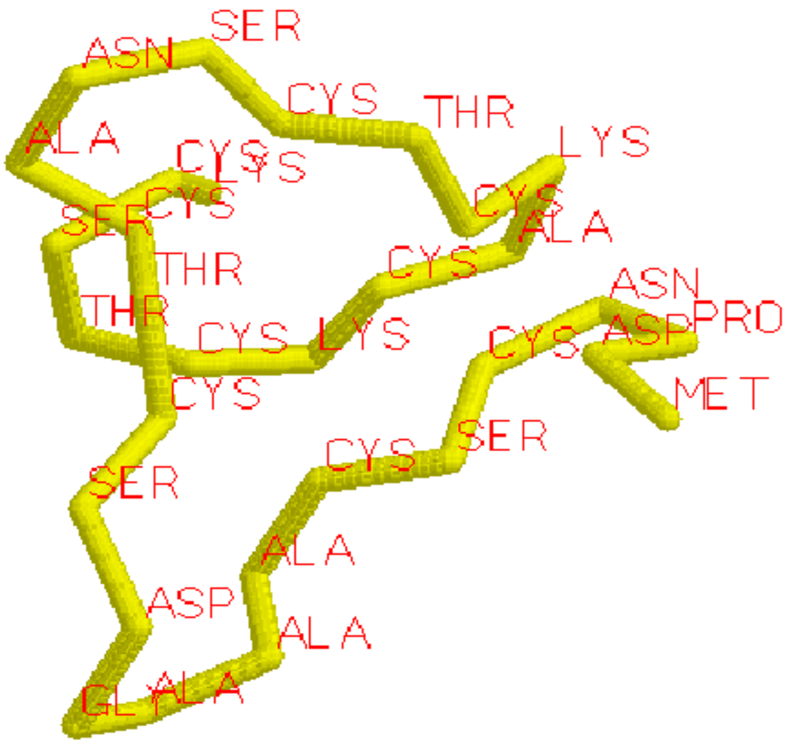
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3057182/

Protein Structure from Computation

- Given an amino acid sequence, e.g.,

MDPNCSCAAAGDSCTCANSCTCLACKCTSCK we can use computation to predict protein folding into a 3D structure.

- Long history: more than 30 years as a “grand challenge” problem in computing
- Useful for Drug design, Enzyme design, Function annotation, Target selection
- On OSG, we presently have two distinct efforts that do this:
  - Baker Lab using Rosetta (see separate talk)
  - F.Zhao using RaptorX (see next slide)

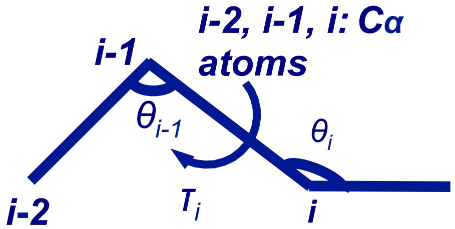


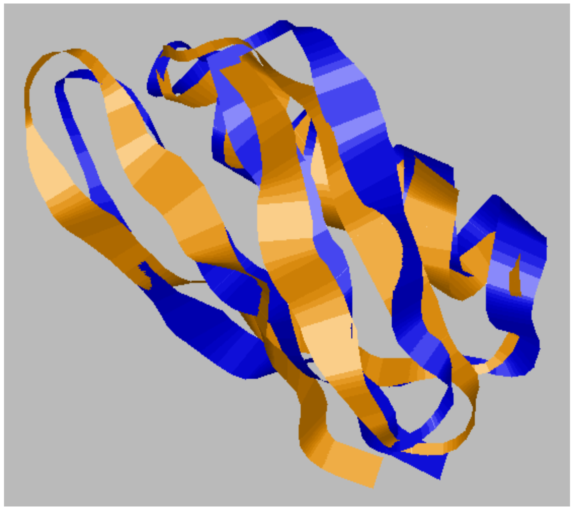
A Computational Solution: CNF-Folder

Sample the 3-D conformation using CNF-Folder

1. Model the relation between the sequence and continuous *Cα*-trace using Conditional Neural Field, a probabilistic graphical model

2. Using the Replica Exchange method to minimize a simple energy function with 2 terms: ***EPAD*** (a novel Evolutionary PAirwise Distance energy) and ***KMBhbond*** (for hydrogen bonds)





One example: 2GB1A.

- Blue is native
- yellow is our prediction

Zhao, F., Li, S., Sterner, B., Xu, J. (2008). Discriminative learning for protein conformation sampling. Proteins.
Zhao, F., Peng, J., Xu, J. (2010). Fragment-free approach to protein folding using conditional neural fields. Bioinformatics

Scientific Community Portals

Examples:

- NanoHub
  - Nano Engineering
- NEESHub
  - Earthquake Engineering
- SBGrid
  - Structural Biology

Commonalities:

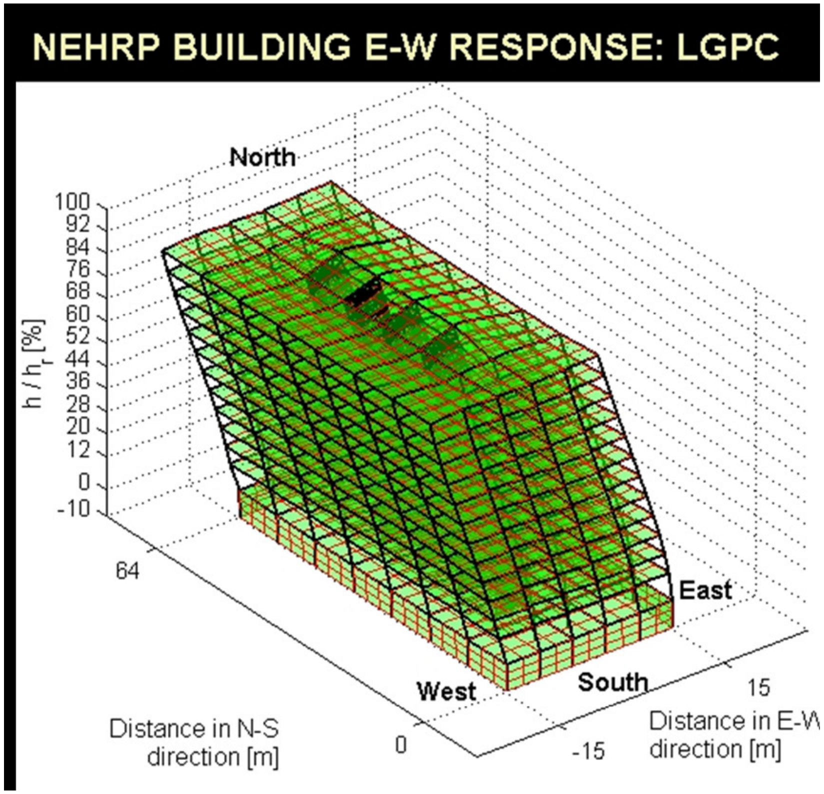
- Support large user communities with diverse skill sets.
- Provide a portal with access to a variety of calculation & simulation tools and data.

OSG’s role:
Facilitate DHTC knowledge exchange. (see backup for SBIR example)
Backend DHTC resources for portals.

Example from NEES

Understand structural integrity of multistory buildings during earthquakes via mix of measurements and simulation.

YouTube Video



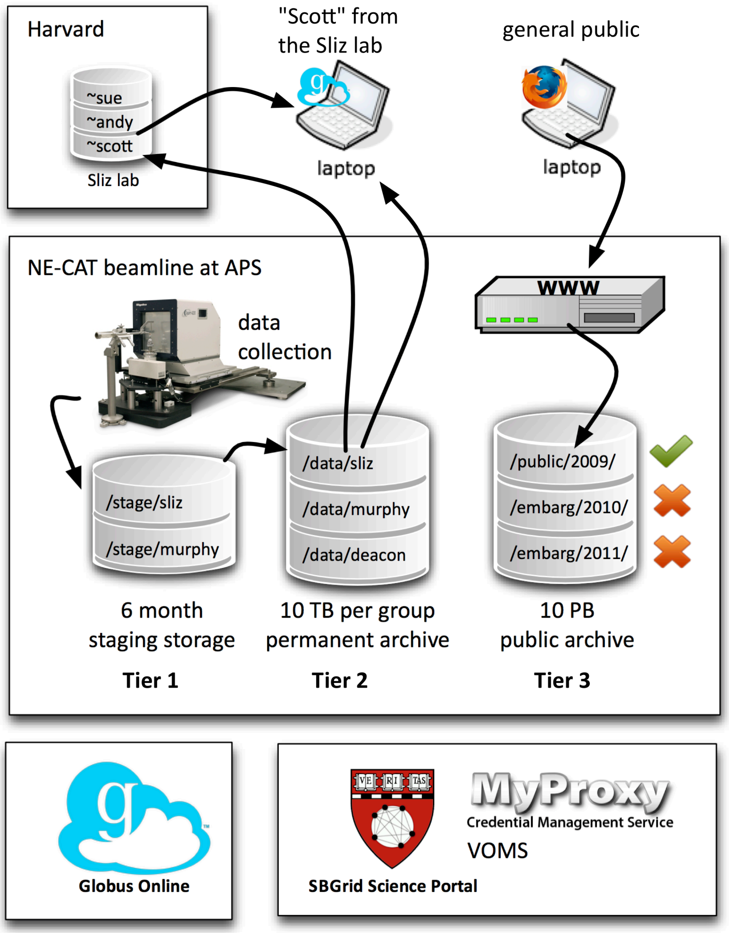
Goal: Improve building codes.

A.R.Barbosa (UCSD), G.Garzoglio, M.Slyz (FNAL)

Example from SBGrid
Data Management: Xray Diffraction Images

Early-stage experimental data in structural biology is generally unmaintained and inaccessible to the public.

SBGrid in collaboration with beamlines from **Argonne National Laboratory** and **Stanford Synchrotron Radiation Light Source** developed a prototype system that adapts existing federated cyberinfrastructure technology to archive and manage diffraction images.



Stokes-Rees, I., Levesque, I., Murphy, F., Yang, W., Deacon, A., and **Sliz, P.** Adapting federated cyberinfrastructure for shared data collection facilities in structural biology. Journal of Synchrotron Radiation. *In press.*

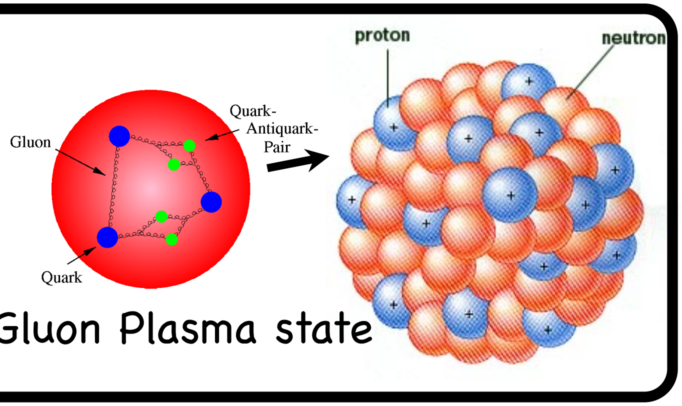
Engage – A special Community

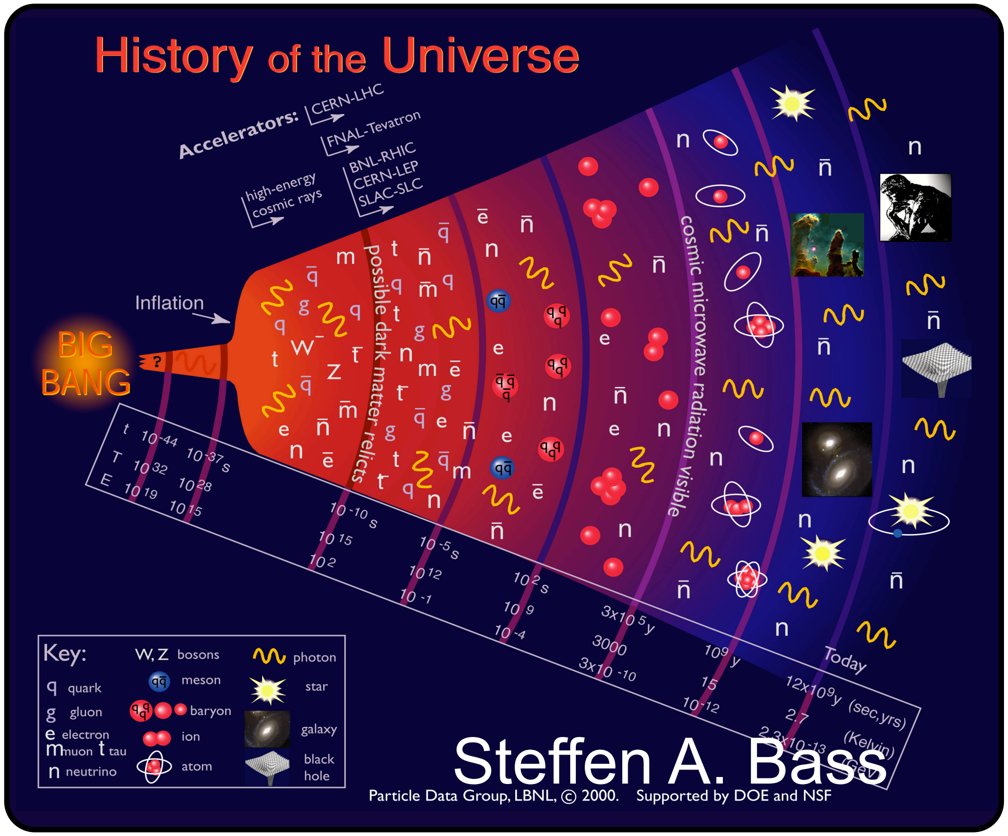
- Incubator for Science Communities on OSG
- Enable single PIs, groups, and new communities to benefit from OSG.
- Operate VOMS for Campus CI
- Examples:
  - – Protein Data Bank (see earlier slide)
  - – Quark Gluon Plasma Simulations (see next slides)
  - – Protein Folding with RaptorX (see earlier slide)
  - – … and much more …

The Big Bang in the Lab: Heavy-Ion Collisions

Quantum-Chromo-Dynamics (QCD):

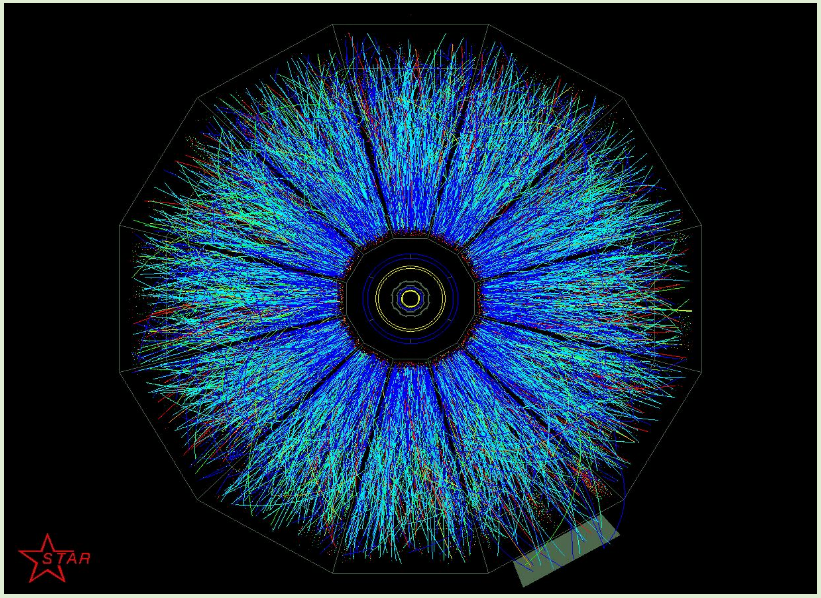
- one of the four basic forces of nature
- holds protons and neutrons together in atomic nuclei
- basic constituents of matter: quarks and gluons
- shortly after Big Bang, entire Universe was in a Quark-Gluon Plasma state





Steffen A. Bass

Particle Data Group, LBNL, © 2000. Supported by DOE and NSF



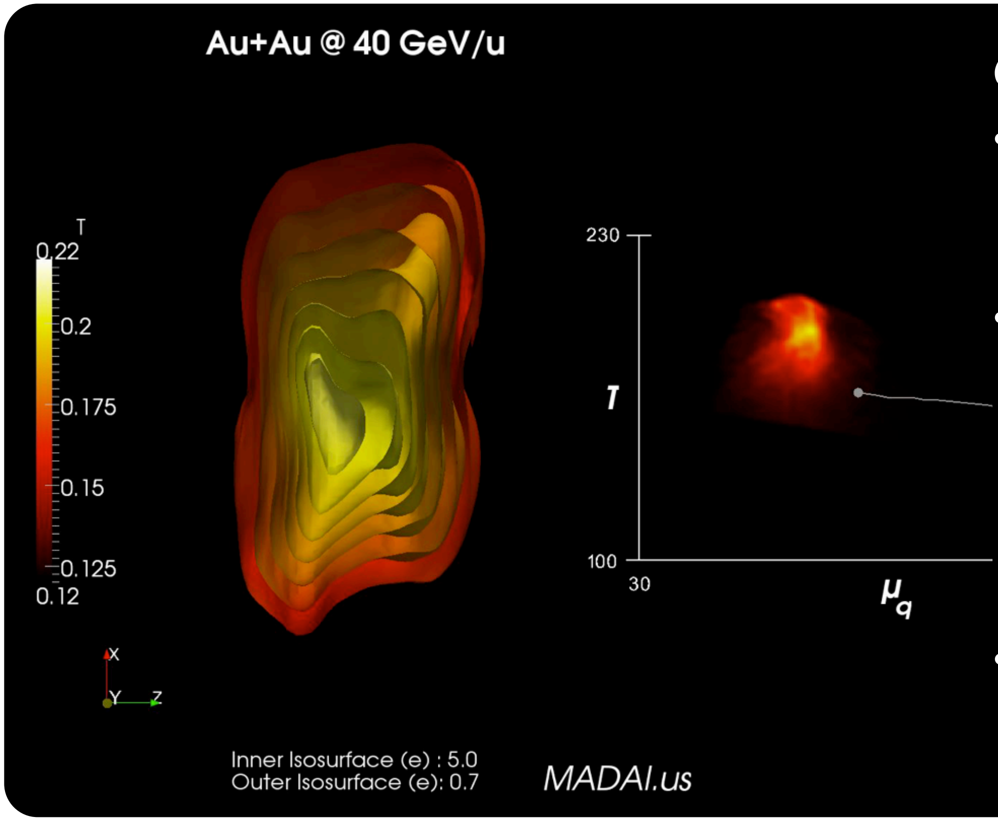
- the QGP can be recreated in collisions of heavy nuclei at particle Colliders such as BNL/RHIC & CERN/LHC

Probing the QGP in Heavy-Ion Collisions

Principal Challenges of Probing the QGP with Heavy-Ion Collisions:

- time-scale of the collision process: 10^{-24} seconds! [too short to resolve]
- characteristic length scale: 10^{-15} meters! [too small to resolve]
- confinement: quarks & gluons form bound states, experiments don't observe them directly

▸ computational models are need to connect the experiments to QGP properties!



Open Science Grid Resources:

- supplied more than 10 million CPU hours over the past 2 years to computational heavy-ion modeling
- made an entirely new class of so-called event-by-event hybrid hydrodynamic + Boltzmann calculations possible, which are to-date the most successful models for simulating QGP creation and evolution
- led to multiple new insights and publications in peer-reviewed journals

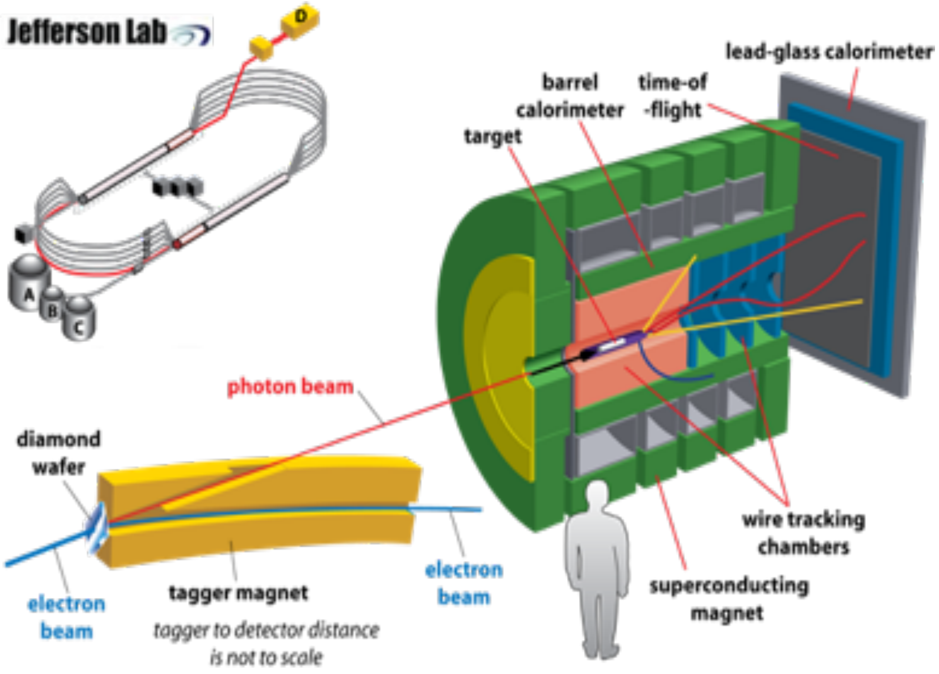
Gluex

Goal: search for explicit gluonic degrees of freedom in the region 1.5 – 2.5 GeV (glueball threshold – charm threshold)

- **Gluex VO** created 9/2009
- Experiment is in *construction phase until 2014*
- **Usage increasing with demand for Monte Carlo**

run period	**usage**
9/2009 – 9/2010	26.4 khr
9/2010 – 9/2011	1.1 Mhr
9/2011 – present	**2.1 Mhr**

- **Plans**: saturate at the level 5-10M cpu.hr/yr until physics data collection begins ca. 2015.
- **Strategy**: glideinWMS – support from OSG admins *outstanding !*



More details in backup

R.Jones

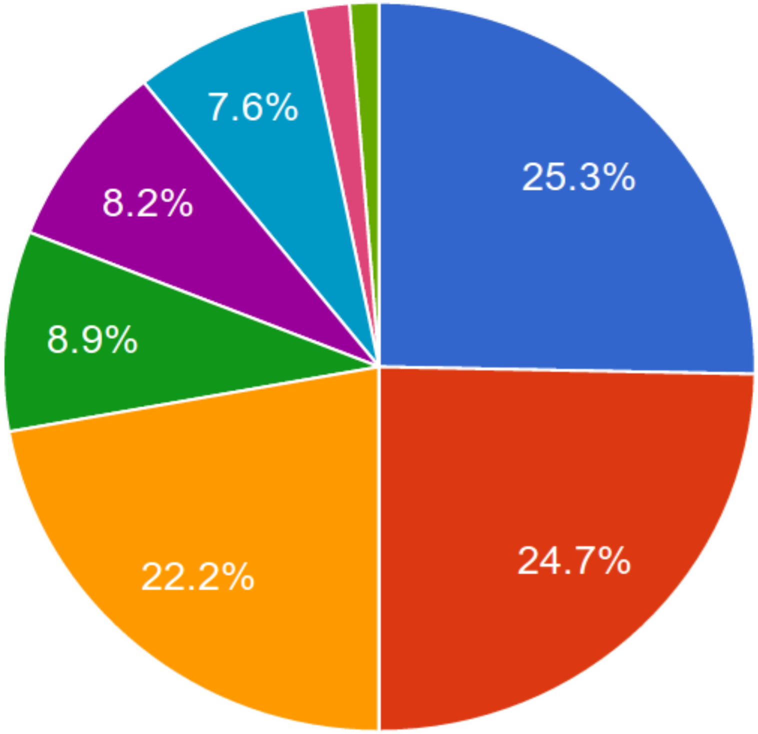
Campus & Regional CI

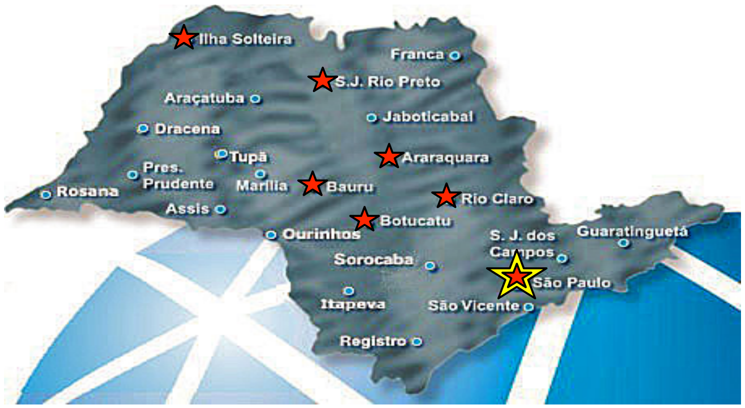
- “local” CI experts bring DHTC solutions to a diverse set of single PIs & small groups.
- Penetration of CI knowledge due to faculty interactions at the local level
- “submit local – compute global” paradigm
  - Scientists learn to operate in local batch environment, then transition seamlessly to (inter-)national CI

Example: GridUNESP

The largest campus Grid infrastructure in Latin America.
Resources dispersed across the state of Sao Paulo (Brazil)

User distribution across Research areas.





See Poster Session and backup material for more details.

Economics on OSG: 2 Examples

Market Demand = $\sum$ Individual Demand

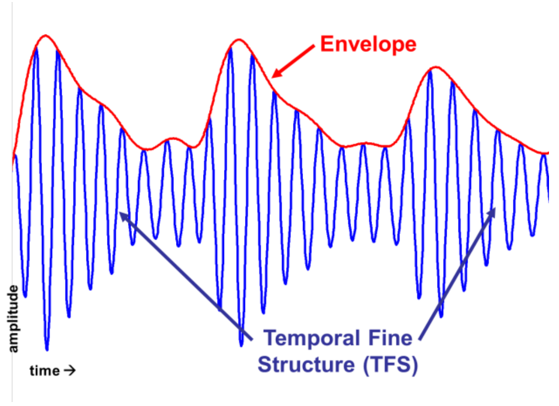
- Must sample individual demand to predict market demand as function of model parameters.
- estimate model parameter by matching observed demands and costs.
- Computational complex.

Study Horse racing as example how heterogeneity in beliefs lead to mispricing in the market.

- Data on 200,000 races
- Compute equilibrium price in each race/market using distribution of beliefs
- Gain insights into the *long shot bias*
  - *Long shots* in horse races are generally overpriced while favorites are underpriced!
- Computationally complex

Amit Gandhi

Better hearing with Cochlear Implants



Normal hearing listeners need sound’s temporal fine structure information to understand speech in noise, localize sounds, and recognize talkers and melodies.

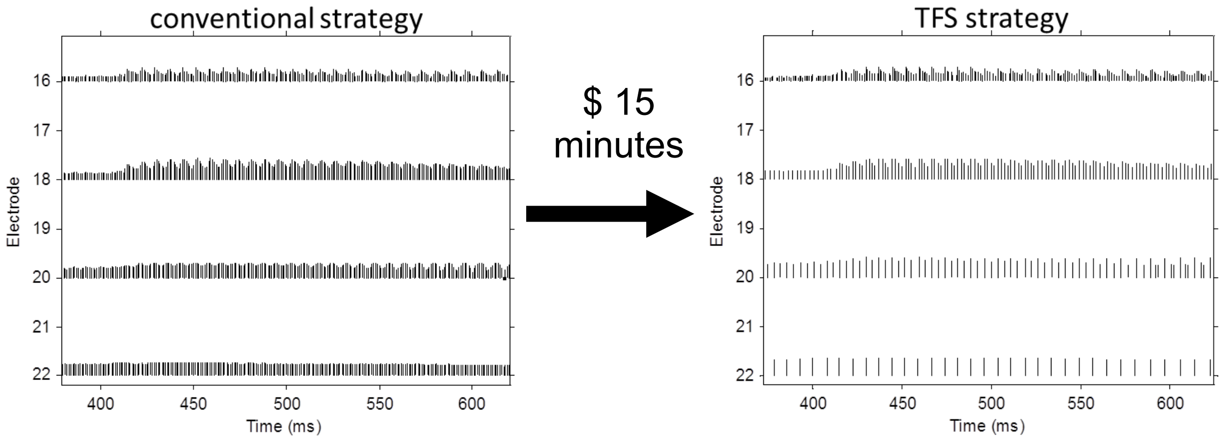
Conventional cochlear implant speech processing strategies discard temporal fine structure and only represent the temporal envelopes of sound.

The algorithm for extracting the fine structure and shifting the pulses is computationally expensive, and creating the over 25,000 stimuli for a cochlear implant experiment would take a lab computer 260 days to complete.

OSG allowed us to make all the stimuli within a day.

Tyler Churchill
Binaurial Hearing and Speech Lab
Waisman Center

Experimental question: If we set the timing for stimulating electric pulses of cochlear implants to represent this temporal fine structure, will cochlear implant listeners hear better?

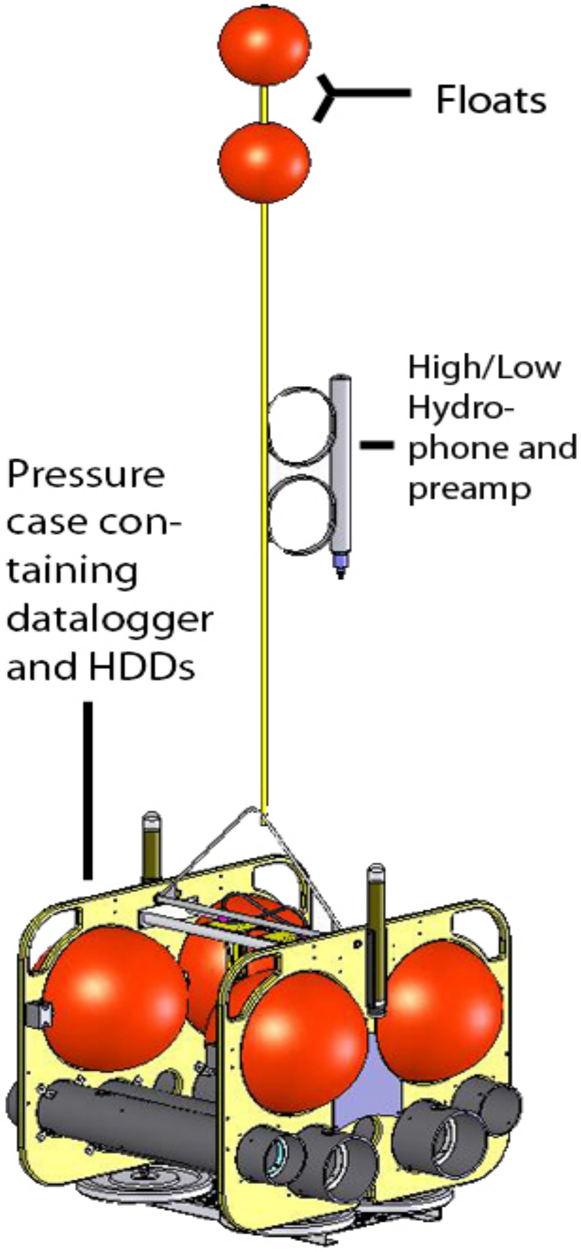


Analysis of WMAP Data

- Wilkinson Microwave Anisotropy Probe (WMAP)
  - Publicly available NASA dataset is analyzed by graduate students from theory and experiment for their thesis projects.
- Topology of the Universe
  - Showed that the Universe is consistent with a torus, ruling out infinite Universe at 4.3 std.dev.
  - ISGTW Article **arXiv:1104.0015**
- Processing WMAP data for gravitational lensing
  - Produce a map of dark matter in the universe.
    - See backup for more detail

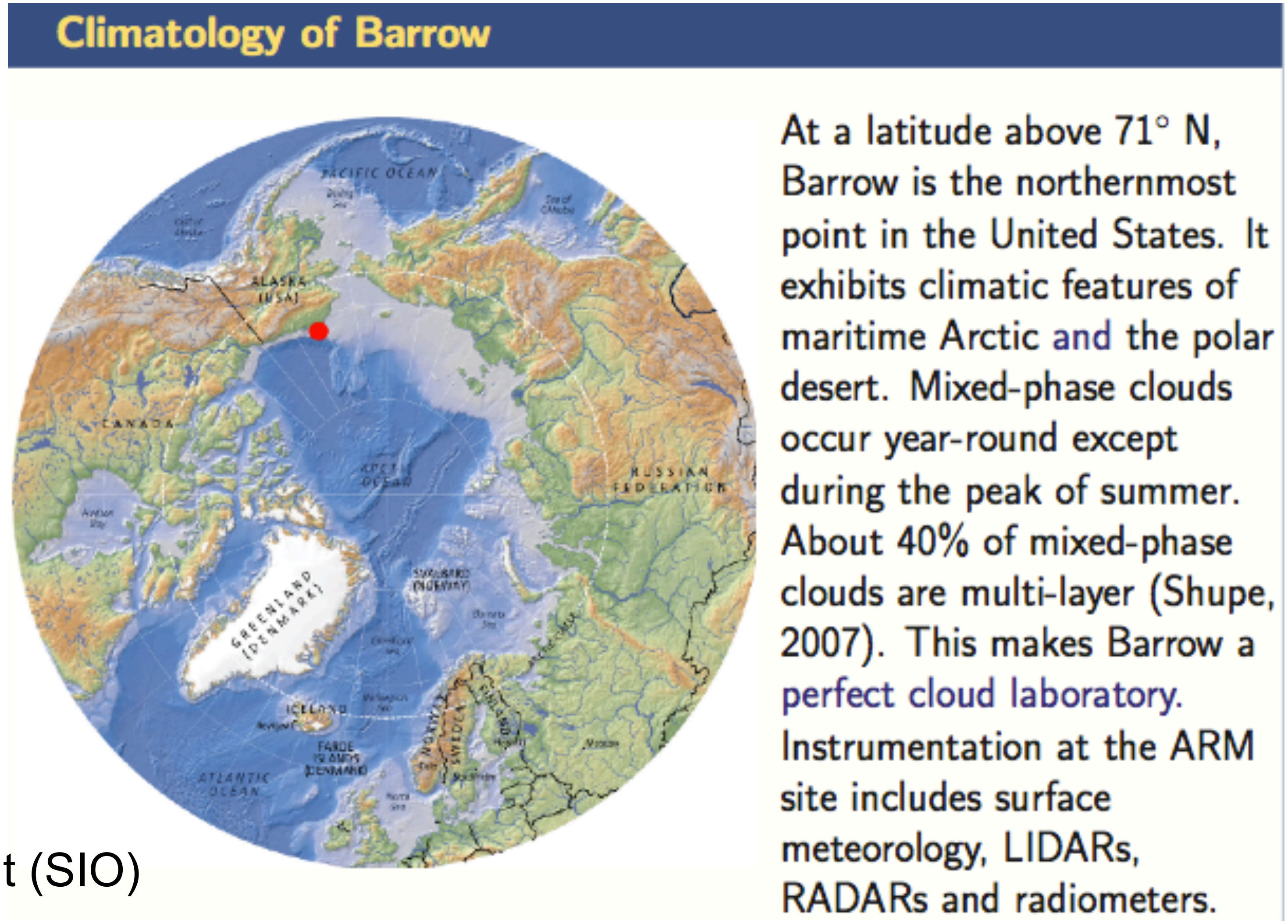
G.Aslanyan, Ch.Feng (UCSD)

Scripps Whale Acoustics Lab on the OSG


Floats
High/Low Hydro-phone and preamp
Pressure case con-taining datalogger and HDDs

- High frequency acoustic recording packages (HARPS) floating in the pacific to acquire acoustic data with up to 320KHz bandwidth for periods as long as a year
- Manual analysis of spectrograms is supplemented by automated detection algorithms that are computationally intensive
- 2 current detection processes run on the OSG:
  - Mid frequency sonar detection to study response of Blue Whales to Anthropogenic Noise
  - Odontocete click detection
- Processing times reduced by an order of magnitude when run in a cluster environment

B.Thayre (SIO)



Climatology of Barrow

At a latitude above 71° N, Barrow is the northernmost point in the United States. It exhibits climatic features of maritime Arctic and the polar desert. Mixed-phase clouds occur year-round except during the peak of summer. About 40% of mixed-phase clouds are multi-layer (Shupe, 2007). This makes Barrow a perfect cloud laboratory. Instrumentation at the ARM site includes surface meteorology, LIDARs, RADARs and radiometers.

J.Muelmenstaedt (SIO)

- Use 11 years of observational data to determine how the meteorological regimes influence cloud formation on the North Slope of Alaska.
- Testing climate models against data for the full annual cycle and substantial climactic variability.

Many Thanks

- To all the people who helped me put this together:
- Mats Rynge, Andreas Prlic, Brooklin Gore, David Swanson, Piotr Sliz, Ian Stokes-Rees, Feng Zhao, Gabriele Garzoglio, Steffen Bass, Richard Jones, Gabriel Winckler, Sergio Novaes, Amit Gandhi, Grigor Aslanyan, Chang Feng, Bruce Thayre, Johannes Muelmenstaedt

Conclusions

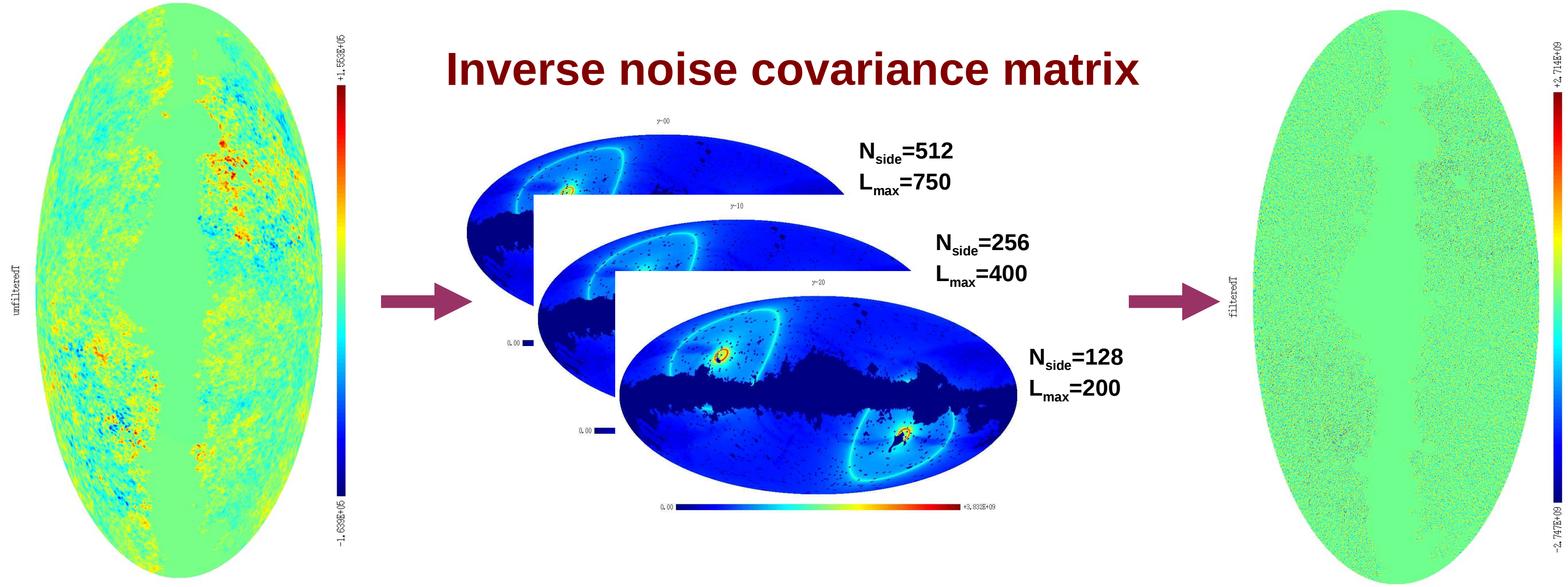
- Protein Structure has become a major Science theme on OSG with multiple very active scientific thrusts at different campuses.
- Science on OSG is very diverse, both in terms scale, and scientific discipline.
- Much of this diversity is due to “word of mouth” at the local level.
- I encourage you to enable your local scientists to submit locally and compute globally.

Apologies

- The choices I made on what to cover in this talk are idiosyncratic.
- No attempt was made to be “inclusive” or “comprehensive”. I chose what I had easy access to.
- Let’s do this again next year, and I promise to pick different but equally idiosyncratic examples.

Backup

Gravitational Lensing Reconstruction WMAP & Gravitational Lensing



Inverse noise covariance matrix

N_{side}=512
L_{max}=750

N_{side}=256
L_{max}=400

N_{side}=128
L_{max}=200

Raw CMB map

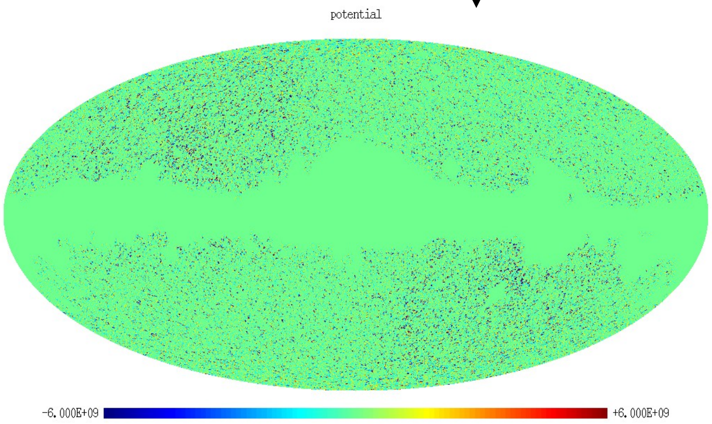
Full covariance-filtered CMB map

multigrid-preconditioned-complex conjugate Gradient (MPCCG)

$$\underbrace{(\mathbf{I}+\mathbf{S}^{\frac{1}{2}}\,\mathbf{N}^{-1}\,\mathbf{S}^{\frac{1}{2}})}_{\text{A}}\,\mathbf{S}^{\frac{1}{2}}\,\mathbf{z}=\mathbf{S}^{\frac{1}{2}}\,\mathbf{N}^{-1}\,\mathbf{m}$$

preconditioner

$$\tilde{\mathbf{A}}=\begin{pmatrix}\mathbf{I}+\mathbf{S}^{\frac{1}{2}}\,\tilde{\mathbf{N}}^{-1}\,\mathbf{S}^{\frac{1}{2}} & 0\\ 0 & \mathrm{diag}(\mathbf{I}+\mathbf{S}^{\frac{1}{2}}\,\tilde{\mathbf{N}}^{-1}\,\mathbf{S}^{\frac{1}{2}})\end{pmatrix}$$



Noisy lensing potential reconstruction from filtered CMB

Bringing new technology to OSG
via a collaboration of an SBIR and a Community Portal

Account Linking for the Grid

Isaac Potoczny-Jones | OSG All Hands Meeting | March 2012

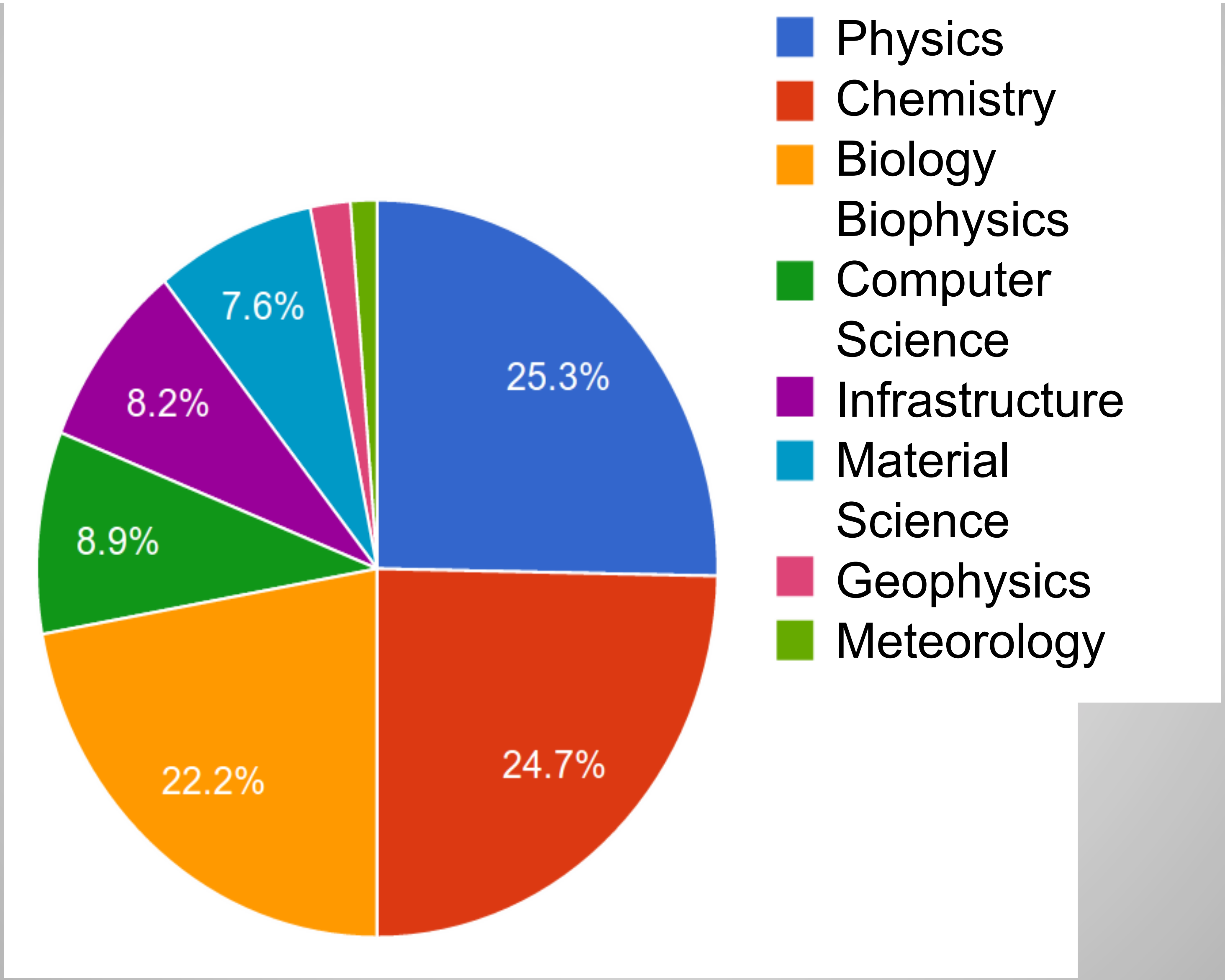
Galois' Account Linking Service (ALS)

- Summary: Helps improve collaboration across organizations
  - By linking multiple user identities together
- Problems addressed
  - Too many credentials & they are hard to use
  - Access depends on which account user logs into
- Approach: Easier to use multiple credentials
  - After linking, users can log in w/ any of their authentication mechanisms
  - Choose the most convenient or most secure (e.g. X.509, passwords)
- Approach: Attribute-based access control across user accounts
  - Users can access data by logging into any of their linked accounts
  - Translates user attributes between e.g. LDAP, VOMS, etc.
- Status: Built as an open-source extension to CAS
  - Initial version complete
  - Trial deployment at SBGrid complete; at US-ATLAS anticipated

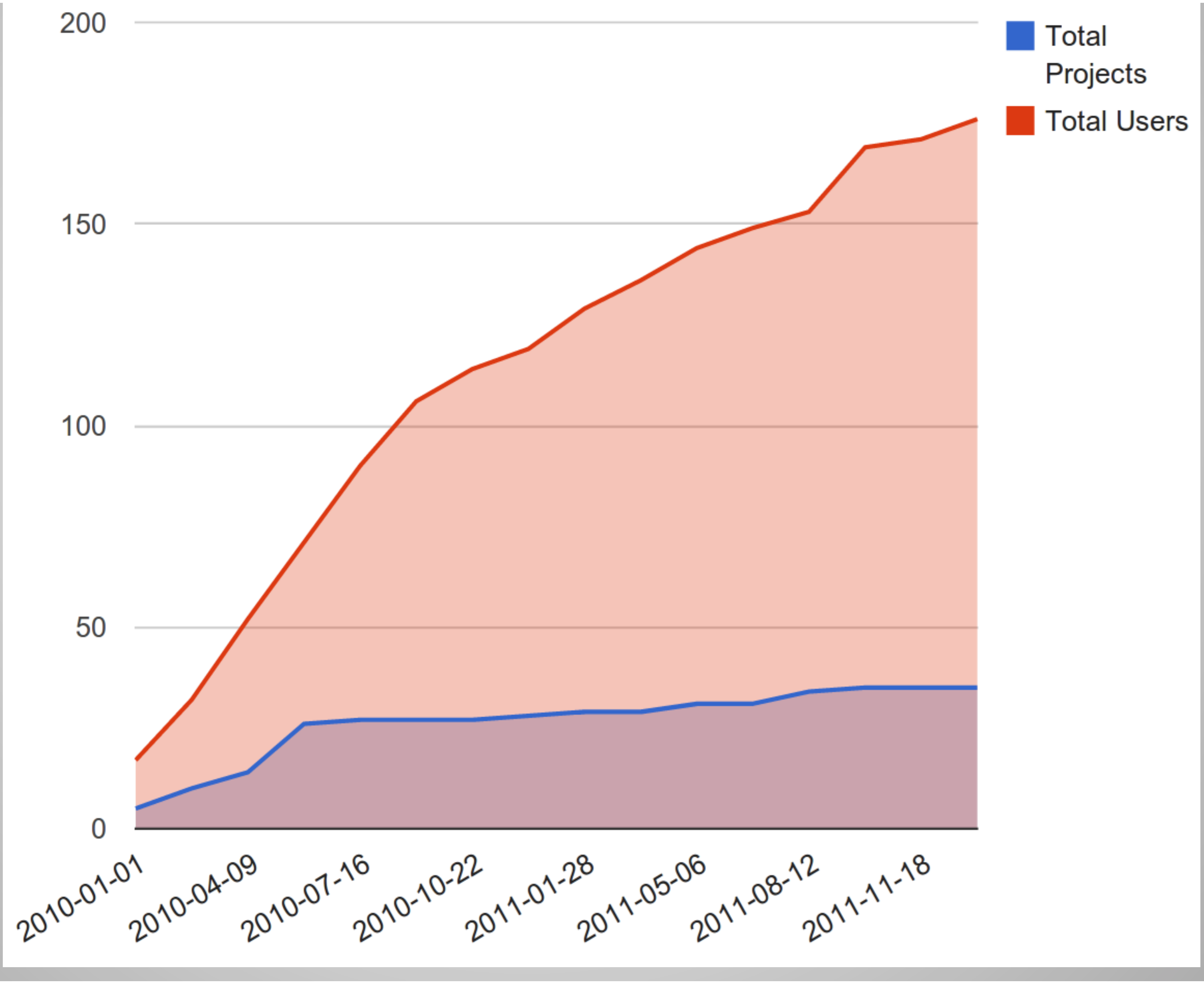
GridUNESP: Research Areas

- Biology and Biophysics
  - Molecular Dynamics
  - Computational Biophysics
  - EEG & Apnea
  - Proteomics
  - Genomics and Phylogenetics
  - Amphibians at high elevations
- Chemistry
  - Modeling for new materials
  - Quantum chemistry
  - Intermetallic phases
  - Coordination compounds
  - Vibrational circular dichroism
- Computer Science
  - Data mining and IPFIX
  - Grid Algorithm optimization
  - Numerical methods
- Geosciences
  - Terrestrial deformation
  - Platform modeling
- Material Science
  - Superconductor vortices
  - Electronic Structure
  - Photo-dissociation of polymers
  - Strong correlated electrons
- Meteorology
  - Historic precipitation in S. Paulo
  - Multi-scale interaction
- Physics
  - Chaos and phase transition
  - Lattice QCD
  - Dark Energy Survey
  - Few-body systems
  - High-energy physics

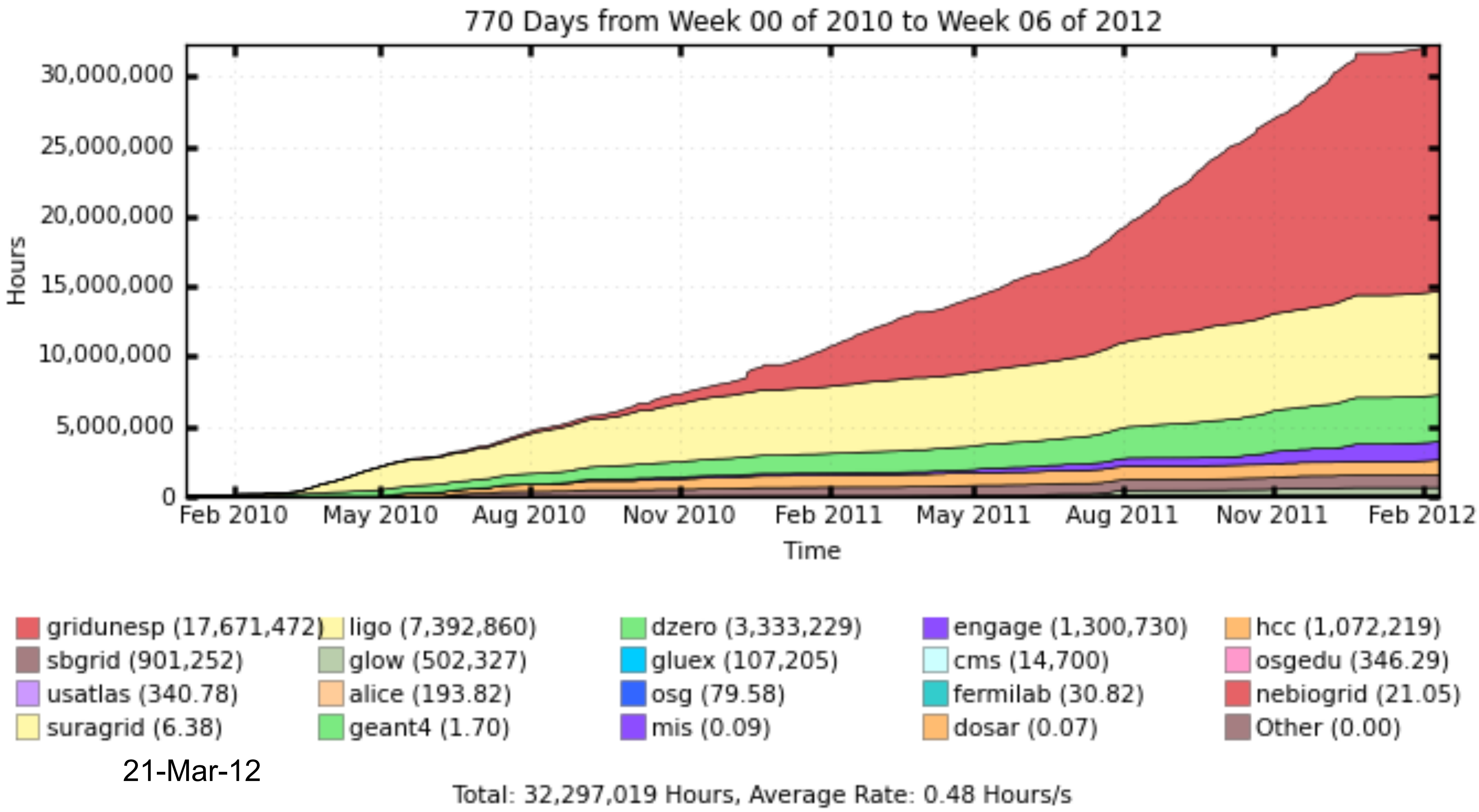
User Distribution per Research Area



Project and User Registrations

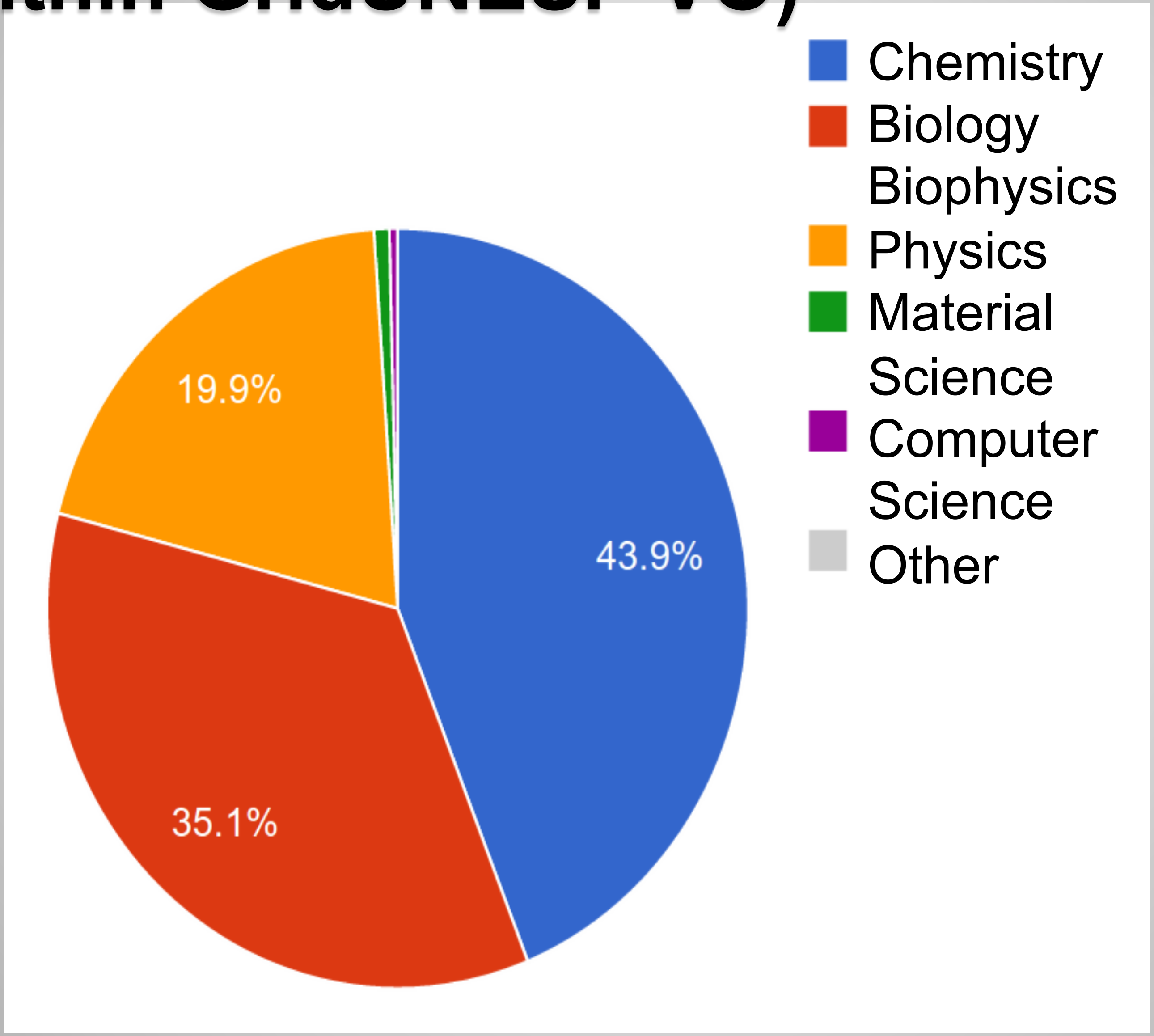


32+ million CPU hours provided to OSG



21-Mar-12

CPU Hours per Research Area (within GridUNESP VO)



OSG Usage by Gluex VO

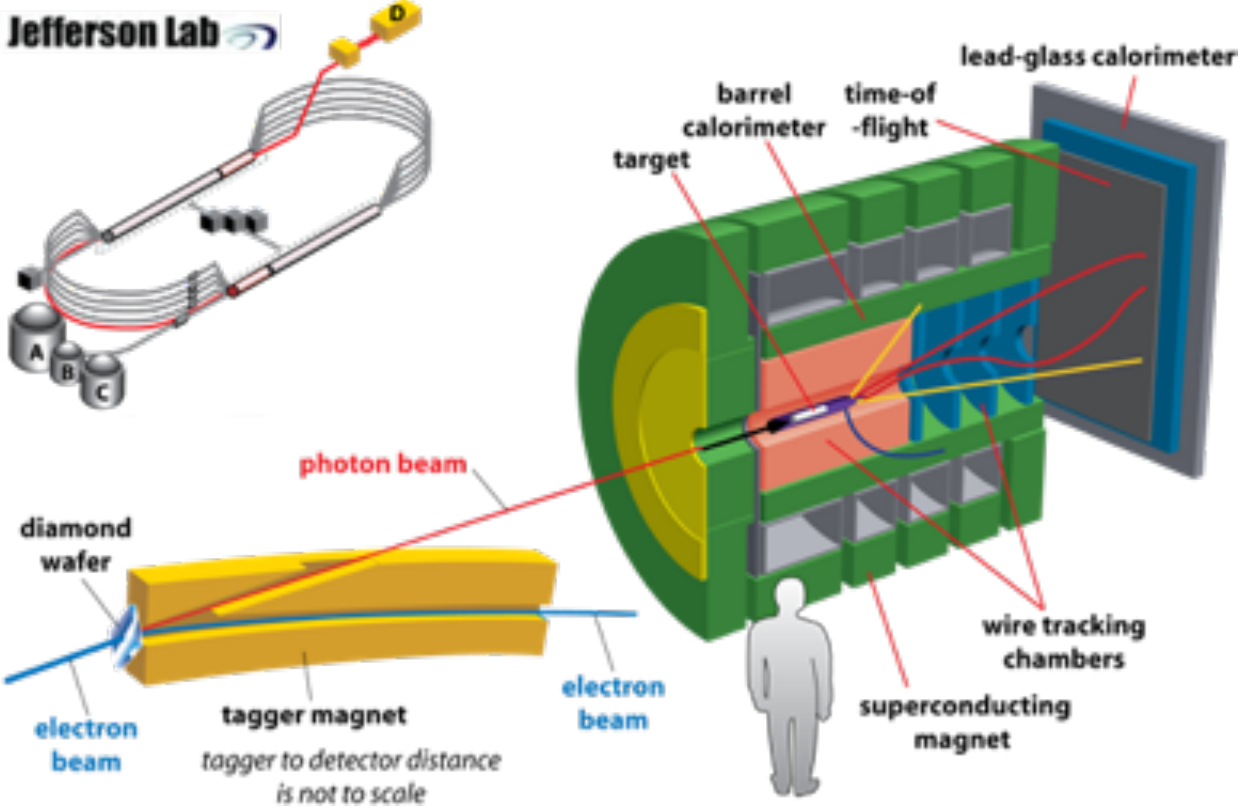
- **Gluex VO** created 9/2009
- Experiment is in *construction phase until 2014*
- **Usage increasing with demand for Monte Carlo**

run period	**usage**
9/2009 – 9/2010	26.4 khr
9/2010 – 9/2011	1.1 Mhr
9/2011 – present	**2.1 Mhr**

- **Growth has slowed as work increases to digest the results**
- **Task:** simulation of background QCD photoproduction (Pythia)
- **Purpose:** develop cuts to suppress background, measure leakage from minimum-bias events into signal sample after cuts, requires very large statistics MC samples, shared between analysis tasks.
- **Plans**: saturate at the level 5-10M cpu.hr/yr until physics data collection begins ca. 2015.
- **Strategy**: glideinWMS – support from OSG admins *outstanding !*

Gluex VO – the science

Goal: search for explicit gluonic degrees of freedom in the region 1.5 – 2.5 GeV (glueball threshold – charm threshold)



Gluex VO – the collaboration

15 institutions + Jlab
~60 members

Collab. Board (6)
Executive Committee
Current spokesperson
Curtis Meyer, CMU

Schedule:

- Sept. 2008: **CD3**
  start of construction
- Dec. 2012:
  end of 6 GeV Ops.
- 2015: **CD4**
  start of operations

Open Science Grid All Hands Meeting,
Lincoln, NB, March 19-22, 2012

- Arizona State University
- Carnegie Mellon University
- Catholic University of America
- Christopher Newport University
- University of Connecticut
- Florida International University
- Florida State University
- University of Glasgow
- Indiana University
- Jefferson Lab
- U. of Massachusetts
- M.I.T.
- North Carolina A&T State
- University of North Carolina
- Santa Maria University
- University of Regina