
Engage on ATLAS
Widening the On-Ramp to Large Scale HTC
Open Science Grid
ISI
renci
RESEARCH ENGAGEMENT INNOVATION

The Engage Virtual Organization

History

Engage has been assisting researchers in migrating science to the High Throughput Computing model via the Open Science Grid since 2007. The community is highly diverse with respect to the kinds of science conducted.

Open Science Grid

Search OSG using Google: search... GO

Search OSG at Work: search... GO

A national, distributed computing grid for data-intensive research.

Home

About the OSG
Learn About Us
What We're Doing
Getting Started with OSG
Contact Us

OSG at work
Calendar of Events
Collaborative Twiki and Chat
Technical Documentation
Security
Software
Education

Information and help
Grid Operations Center
Documentation and Reference
Monitoring

Can't Find What You're Looking For?

Home > Protein structure-taking it to the bank

Protein structure: taking it to the bank

Properly functioning proteins are essential for our bodies. A protein's structure, the folded form its amino acid string assumes, determines its function. Scientists know the sequence and structure of about 50,000 proteins—out of millions. They keep this valuable information in a "bank"—the NSF- and NIH-funded Protein Data Bank. To predict the structure of a newly identified protein, scientists can compare it to a similar banked protein.

This works quite well, but what if there is no similar protein in the bank? Then it's back to first principles: creating predictions from scratch using the physical principles that describe the interactive forces between atoms. Experimental methods, such as x-ray crystallography and nuclear magnetic resonance are time-consuming and costly, and cannot be used on all proteins, so scientists have turned to computational predictions.

At left, the predicted (gold) and experimental (blue) structure of 30S ribosomal protein S27A from Thermoplasma acidophilum.

At right, the same for protein G.

Xu's team participated in the 2008 Critical Assessment of Structure Prediction (CASP) competition, sponsored by NIH, BioSapiens Network and the European Molecular Biology Organization, in which participants are given 120 proteins with unknown structures to predict. RAPTOR ranked among the top 5 of 85 teams. Thermoplasma acidophilum, represented above, was a test protein used in CASP.

Image courtesy of Jinbo Xu.

RAPTOR is short for Rapid Protein Threading Predictor

Image courtesy of bioinformaticssolutions.com.

Researchers at the Toyota Technological Institute at Chicago, an affiliation of the University of Chicago, are using the Open Science Grid to quickly and accurately simulate protein structures, and thereby to determine their functions.

An average protein contains about 250 amino acids, each of which has at least 10 atoms. To develop a prediction, scientists must determine the position of the atoms in the system. Each atom has many possible positions. Scientists must search all possible conformations to find the most stable one—a computationally demanding process, says Toyota Technological Institute researcher Jinbo Xu.

RAPTOR on OSG: quick and accurate

Xu and his team use RAPTOR, a molecular modeling software package developed by Xu, to run thousands of "small protein" (under 100 amino acids) simulations on OSG. RAPTOR, available to any researcher, samples all possible configurations against their mathematically estimated probability of being stable. They evaluate the results to find the most stable, hence the most likely to be the "true" structure. Xu's team assures RAPTOR's accuracy by predicting experimentally determined protein structures and comparing the predictions with the known structures.

The simulations are independent of each other, so OSG's distributed framework is ideal for running them quickly and simultaneously.

"OSG has helped us shorten our folding simulation experiments from months to days and is now an essential computing platform for our research," Xu says.

The Engage Virtual Organization

Goals

- Help researchers execute computationally intensive science at large scales.
- Enable access to High Throughput Computing resources.
- Provide hosted HTC tools and services.
- Investigate emerging usage paradigms including multi-core architectures.

The Engage Virtual Organization

Approach

- Establish direct relationships with researchers.
- Work with a highly diverse community.
- Consult directly on application requirements.
- Provide design, development, monitoring and assessment assistance.
- Enhance infrastructure iteratively in response to the evolution of the community.

The Engage Virtual Organization

Infrastructure

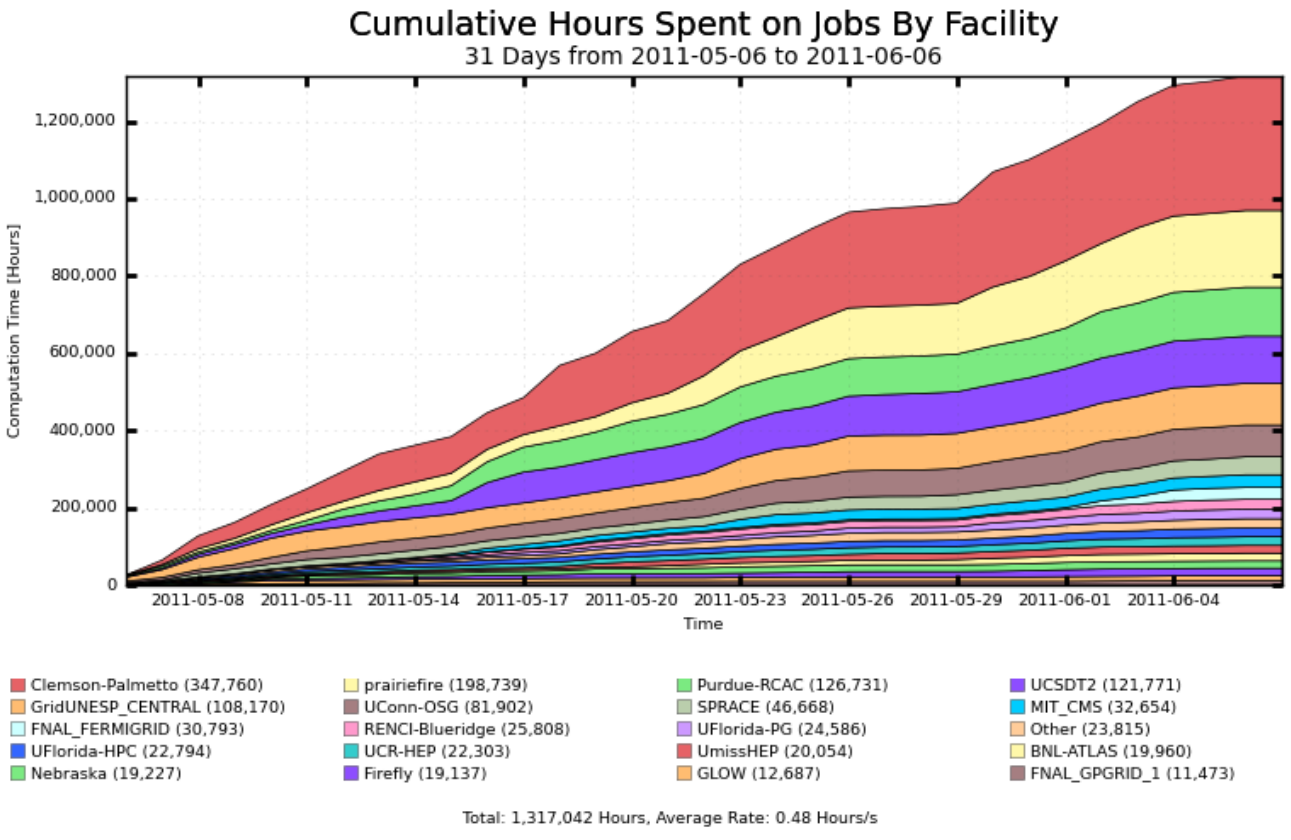
- Pilot Submission: **GlideinWMS**
- GlideinWMS Factory: **UCSD**
- Workflow Management: **Pegasus**
- Architecture specific groups (Bigmem, HTPC)
- Expertise in MPI, GPGPU and Pegasus
- Large data is pre-staged where possible

The Engage Virtual Organization

Infrastructure

Sites Include:

- Nebraska HCC
- Clemson
- Uconn
- Umiss
- Fermilab
- RENCI
- Purdue
- Uflorida
- UCSDT2
- MIT_CMS
- SPRACE



The Engage Virtual Organization

Approach - Detail

- Workflows developed on local clusters.
- Engage CI-Team specialists map to OSG.
- Standard scripts used to execute on OSG.
- Jobs target appropriate sites via GWMS groups.
- Expertise in MPI, GPGPU and Pegasus

The Engage Virtual Organization

Applications

Type	Example	Description
Standard	**Hannah Petersen** Memory: 500MB Wall Time: 6-8hrs	File transfer: globus-url-copy No external dependencies (eg. Java, etc.)
Bigmem	**Andreas Prlic** Memory: 2GB Wall Time: 14 hrs	Protein Data Bank (PDB) pre-staged File transfer: globus-url-copy Requires Java 1.6 (pre-staged)
HTPC	**Laura Perissinotti** MPI: 8 CPU Wall Time: 8 hrs	Uses High Throughput Parallel Computing model. 8-way jobs delivered via OSG, communicating via shared memory.
Blastp	**Juan Antonio Raygoza** Memory: <1.7GB Wall Time: 20 hrs	Data: Uniprot & Fasta databases.

The Engage Virtual Organization

General Engage Requirements

- Worker Node OS = RHEL 5
  - CentOS5, SL5 OK!
  - We run/test on CentOS 5.
- Worker nodes with internet access
  - Can be NAT.
  - Cannot just be http proxy.
- OSG Certs
  - Required for GlideinWMS

The Engage Virtual Organization

General Engage Requirements

- OSG WN-Client
  - Should already be everywhere
  - We use globus-url-copy and wget
- Worker node local disk (flexible)
  - Assume 5GB per slot
  - In practice 500MB will work.

The Engage Virtual Organization

General Engage Requirements

- Worker node memory (flexible)
  - Assume 1.5-2GB per slot
  - 1 GB minimum
  - We detect memory with GlideinWMS
  - We use a special Bigmem group for >2GB
- Squid Cache (Optional)
  - Or any http proxy
  - We hope to migrate to using it for data 10MB < X < 150MB
- We're fine with preemption.

Engage Blog

See the blog for

- Information on the submit host
- Descriptions of Engage jobs
- Discussion of older submit infrastructure