

# Collaborative Use Example: WeNMR/SBGrid

## Collaborating teams in Europe and the US

**WeNMR** ([www.wenmr.eu](http://www.wenmr.eu)) is both a three years project funded under the European Commission's 7th Framework Programme (e-Infrastructure RI-261571) and a Virtual Research Community supported by **EGI**, the largest one within the life sciences, with over 500 VO members. WeNMR aims to bring together complementary research teams in the structural biology and life sciences arenas to form a worldwide virtual research community and provide them with a platform integrating and streamlining the computational approaches necessary for NMR and SAXS data analysis and structural modelling.

**SBGrid** ([www.sbgrid.org](http://www.sbgrid.org)) is a global consortium of 220 structural biology groups and an Open Science Grid Virtual Organization. SBGrid laboratories share a common collection of approximately 350 scientific applications compiled and optimized for Linux and OSX operating systems. To support a subset of CPU-intensive workflows, SBGrid also runs a scientific portal, which interconnects with the computational resources of the Open Science Grid (developed with support from NSF and OSG and open to all members of the structural biology community). SBGrid operates out of the Department of Biological Chemistry and Molecular Pharmacology at Harvard Medical School.

**The collaboration** between the WeNMR project and SBGrid dates back to the previous e-NMR project (2007-2010) and has been well established since the start of WeNMR (2010-2013). In a joint SBGrid/OSG – WeNMR/EGI meeting held in March 2011, we decided to implement interoperability between EGI and OSG in the domain of Structural Biology. A number of key persons were assigned to each project to follow a well-defined work plan that was implemented in the following months. The work was presented at the EGI Community Forum in March 2012, and is published in section 3.2.1 of [http://pos.sissa.it/archive/conferences/162/040/EGICF12-EMITC2\\_040.pdf](http://pos.sissa.it/archive/conferences/162/040/EGICF12-EMITC2_040.pdf).

## Scientific justification for the collaborative activity

WeNMR and SBGrid developed two complementary grid portals that are used by structural biologists globally to perform specialized, compute intensive structure determination workflows. The WeNMR portal supports a subset of services that are used for NMR structure determination (computations from data processing to structure calculations) as well as for molecular modelling (e.g molecular dynamics or macromolecular docking), while SBGrid focuses on advanced X-ray crystallography structure determination workflows (e.g. wide search molecular replacement or low-resolution structure refinement). The two are thus highly complementary. By coordinating the efforts of both projects, we aim to further expand the scope of services and continue to make them available to the structural biology community at large.

## Justification for access to resources in both the US and Europe, where applicable

The two portal infrastructures developed by WeNMR and SBGrid are unique in their design and require a high level of computational and scientific support. These infrastructures need to be continuously updated to incorporate new versions of scientific applications and databases. It is, therefore, very appealing to extend access to those specialized resources to the global community rather than to redeploy or redevelop supported workflows in parallel in the USA and Europe. Sharing access to global resources would have the additional advantage of load balancing across a larger pool of computational resources and of sharing and promoting specialized local knowledge and technologies.

## Resource levels required to support the collaborative work.

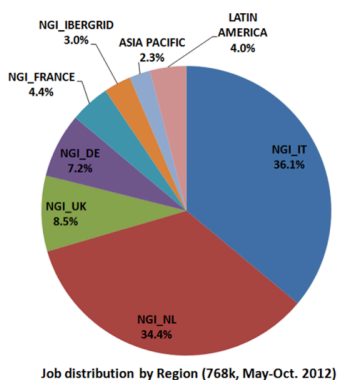


Figure 1. WeNMR Resource Centers

On an annual basis, the WeNMR project supports 4,000,000 CPU hours with peak utilization of ~5,000 CPUs. WeNMR submits jobs to 25 mostly European resource centers (RC) and has opportunistic access to ~50,000 CPU-cores (**Figure 1**). Only 3 out of 25 RCs are fully dedicated to the WeNMR project and provide a high-availability 320 CPU-core queue. The SBGrid VO supports approximately 5,000,000 CPU hours on the Open Science Grid, with peak utilization of ~5,000 CPUs. SBGrid can submit computational jobs to 20 resource centers in the US and has opportunistic access to approximately 40,000 CPU-cores. Between 100 and 10,000 cores are typically available to SBGrid for immediate job execution.

Both VOs attempt to load balance computations submitted by the end users, however the rate of job submission is variable and the convergence of the majority of computations can rarely be predicted based on job submission parameters. Resource utilization is, consequently, intermittent. The periods of peak utilization between the two VOs usually do not overlap and therefore access to a combined

pool of resources between EGI and OSG would effectively double the computational capacity of both VOs.

## Patterns of use

Several structural biology applications provided by the WeNMR Science Gateway are very popular in the US, as shown in the map below for the HADDOCK application. A large fraction of the 2,800+ HADDOCK users are located in the US (**Figure 2 top**). The WeNMR CS-Rosetta portal, which is responsible for a large fraction of CPU usage on the grid (>80%), is also utilized extensively by US users (~14% of registered users). Correspondingly, the majority of users who access SBGrid WSMR and low-resolution refinement portals are from the US and Europe. In addition, statistics generated by WSMR computation provide feedback to a software development project in the UK and a large fraction of SBGrid consortium members are in European laboratories (**Figure 2 bottom**).

Grid jobs submitted through the WeNMR portal are usually dispatched to the EGI resources (**Figure 2**). Personal X.509 certificates for US users who require access to WeNMR resources are issued by the SBGrid VO. WeNMR computations are typically completed within one to two days for HADDOCK, and one to two weeks for CS-Rosetta. The number of concurrent runs per user is, however, limited. SBGrid computations are dispatched to the OSG resources. In most cases all large computational jobs submitted through the SBGrid portal are completed within 1-2 weeks, although occasional bottlenecks can develop with wait times of up to four weeks.

**A collaborative pilot project** was established in 2011 to prototype a framework that would be required for full interoperability between the two projects (**Figure 3**). The plan was carried out in two phases. In Phase I, the proof of concept that WeNMR application jobs can effectively run over OSG resources was obtained by using EMI/gLite grid services set up in Europe and appositely enabled with the SBGrid VO.

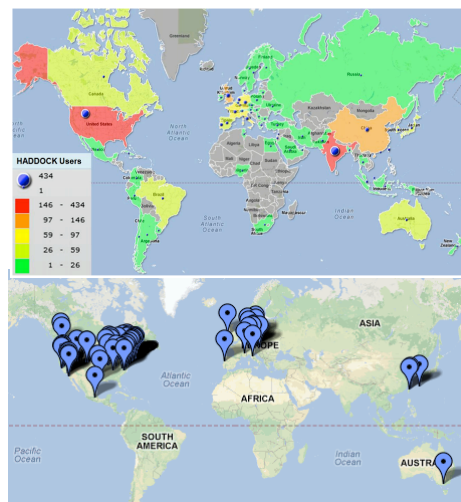
In Phase II, a Condor submit node was installed in the same server hosting the gLite-UI at INFN-Padova and configured to work with the GlideinWMS Frontend node hosted at Harvard Medical School (HMS) in Boston and operated by the SBGrid VO. The Frontend interacts with the Factory node hosted at the University of California, San Diego (UCSD) and operated by OSG. While glideins are submitted by the Factory to the OSG nodes using an SBGrid VO proxy, WeNMR jobs can be submitted to the Condor pool using an enmr.eu VO proxy, as done when submitting to EGI via the gLite-WMS. The WeNMR application software was subsequently installed on the OSG CEs and CS-Rosetta jobs were submitted to test the system.

The current pilot system developed by WeNMR and SBGrid supports grid interoperability while allowing grid operators to control access and utilization. The developed system can provide US scientists who utilize the European portal infrastructure with opportunistic access to the NSF and DoE sponsored computational resources. Interoperability and collaboration will be crucial in the evolving e-Infrastructure landscape, in particular for the long-term sustainability of the services provided: for both sides, giving computational access to users outside their respective geographic/political locations (North America, EU) can only be justified provided that a proper mechanism for sharing of resources is in place.

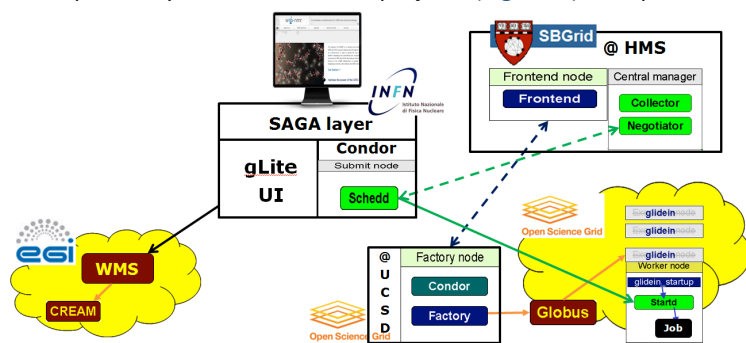
## Limiting Factors

Further resources are required to fully implement the interoperability between two systems:

- **A pilot project to submit jobs from WeNMR to SBGrid/OSG needs to be transitioned to production.** This project will require significant work to adapt the job management scripts lying behind the WeNMR portals, developed in the past according to the EMI/gLite environment, to the new Condor environment, or alternatively, to implement a SAGA layer to decouple the application level from the underlying middleware stack.
- **A reciprocal system allowing submission of SBGrid jobs to EGI needs to be developed and implemented.**
- **A more flexible job/CPU accounting system will need to be put in place.** This system will provide a basis for load balancing between two VOs based on cumulative usage rather than dispatching individual computations (e.g. WeNMR VO will count the total number of CPU-hours dedicated to support of US research laboratories, whether those jobs are computed on EGI or OSG resources, and then “claim” those hours from OSG when an EGI resource contention develops).



**Figure 2: Top) Global distribution of WeNMR HADDOCK users. Bottom) Structural biology laboratories utilizing SBGrid software support.**



**Figure 3. WeNMR-SBGrid pilot project to establish EGI-OSG interoperability.**