



University of Virginia: Protein Folding on the Grid

NSF Middleware Initiative (NMI) Integration Testbed Case Study Series

Series contact: Mary Fran Yafchak, Southeastern Universities Research Association,
maryfran@sura.org.

The NMI Integration Testbed Program provided practical evaluation of NMI components within the context of real projects and application scenarios from June 2002 through November 2004. During that time, NMI Testbed sites collectively submitted over 220 evaluation reports to middleware component developers as direct feedback into the NMI development cycle. Site representatives also actively inspired, promoted and facilitated the integration of middleware throughout their institutions.

The NMI Integration Testbed Case Study Series documents the most significant influences and outcomes of NMI Testbed sites' middleware integration efforts, highlighting intersections with established projects, application contexts and influences, drivers for innovation, decision points and challenges. Through this documentation, the work of these pioneering institutions is captured to provide a breadth of insight and approaches for others to use towards successful middleware development and deployment.

This NMI Integration Testbed Case Study Series is sponsored by the National Science Foundation Middleware Initiative-Enterprise and Desktop Integration Technologies (NMI-EDIT) Consortium of EDUCAUSE, Internet2, and SURA. Additional support was provided by the National Science Foundation Cooperative Agreement NSF 02-028, ANI-0123937.

Copyright © 2004 University of Virginia. The University of Virginia permits use of this content for noncommercial purposes with proper attribution. All rights reserved.



Executive Summary

Researchers in the field of biophysics study the nature of the molecules, such as proteins, from which living organisms are composed. Their work helps to find cures for a number of human diseases caused by problems with specific proteins. The process by which proteins are created is called “protein folding”. In order to prevent and correct protein folding problems, scientists must have a thorough understanding of the specifics of the folding process for each type of protein.

The investigative work of biophysics researchers is highly computationally intensive. Their work has often been throttled back due to a lack of sufficient computing resources. Fortunately, grid computing can help to overcome some of these resource limitations.

Marty Humphrey, Assistant Professor in the Computer Science Department, is a long-standing leader of the University of Virginia’s (UVA) work in grid computing, and in bringing the benefits of the grid to biophysics researchers on his campus. Humphrey’s team has significantly enhanced the computing environment for UVA’s researchers through the installation of a Computational Biophysics portal. While the base functionality of grid software provides the user with a layer of “insulation” from the complexity of this environment, portals

provide an even more seamless, user-friendly computing interface.

UVA’s portal work was done in conjunction with their membership in the National Science Foundation’s (NSF’s) National Partnership for Advanced Computational Infrastructure (NPACI), and it also coincided with their participation in the National Science Foundation Middleware Initiative (NMI) Integration Testbed¹. There are key technologies from the GRIDS Center utilized for UVA’s Computational Biophysics portal and the NPACI grid. UVA found that the NMI package pieces fit nicely together, with minimal additions (other than application-specific) when used with applications such as CHARMM and Amber.

In addition to Humphrey’s grid work, UVA’s Jim Jokl is a leader in security and interoperability issues associated with the public key infrastructure (PKI). By participating in the NMI Testbed, Humphrey and Jokl were able to more closely consider the integration of the Computational Biophysics portal with the campus infrastructure, particularly authentication

¹ As part of its overall effort to develop and disseminate software that lets scientists and educators share resources across the Internet, NMI has begun a practical deployment and evaluation effort called the NMI Integration Testbed. Managed by the Southeastern Universities Research Association (SURA) on behalf of the NMI-EDIT Consortium (NSF Middleware Initiative-Enterprise and Desktop Integration Technologies; <http://www.nmi-edit.org/>), the testbed consisted of eight universities that participated in a closely coordinated effort to deploy and evaluate NMI technologies. <http://www1.sura.org/3000/NMI-Testbed.html>



infrastructures. A significant lesson to come out of Humphrey and Jokl's work is that when people set about deploying NMI on their campuses, they should attempt as much as possible to reuse existing authentication infrastructures in their grid deployments.

For more information about UVa Protein Folding on the Grid, contact Marty Humphrey at humphrey@cs.virginia.edu.



NMI Components Highlighted in this Case Study

The NMI components discussed in this case study series encompass NMI Releases 1 through 4. Information about NMI Releases can be found at <http://nsf-middleware.org/>.

Condor-G

The GRIDS Center's Condor-G is a computation management agent for the grid. Condor-G is the marriage of technologies from the Condor project and the Globus project (see below).
Home site: <http://www.cs.wisc.edu/condor/>; Globus (see below).

Globus

The GRIDS Center's Globus Toolkit is an open-source collection of modular technologies that simplifies collaboration across dynamic, multi-institutional virtual organizations. It includes tools for authentication, scheduling, file transfer and resource description.
Home site: <http://www-unix.globus.org/toolkit/>

GridPort

The GRIDS Center's GridPort enables the development of portals and applications on top of underlying distributed and grid computing infrastructure to facilitate computational science.
Home site: <http://gridport.net/index.cgi>

GSI OpenSSH

The GRIDS Center's GSI-OpenSSH is a modified version of OpenSSH that adds support for GSI authentication, providing a single sign-on remote login capability for the grid. GSI-OpenSSH can be used to login to remote systems and transfer files between systems without entering a password, relying instead on a valid GSI credential for operations requiring authentication.
Home site: <http://grid.ncsa.uiuc.edu/ssh/>

MyProxy

The GRIDS Center's MyProxy is a credential repository for the grid. MyProxy provides a set of flexible authorization mechanisms for controlling access to the repository.
Home site: <http://grid.ncsa.uiuc.edu/myproxy/>

MPICH-G2

The GRIDS Center's MPICH-G2 is a grid-enabled implementation of the MPI v1.1 standard based on the popular MPICH library developed at Argonne National Laboratory. That is, using services from the Globus Toolkit(R) (e.g., job startup, security), MPICH-G2 allows you to couple multiple machines, potentially of different architectures, to run MPI applications.
Home site: <http://www3.niu.edu/mpi/>

Shibboleth

The Shibboleth technology supports inter-institutional sharing of web-based resources subject to access controls.
Home site: <http://shibboleth.internet2.edu>



University of Virginia: Protein Folding on the Grid

There are a number of human diseases that are caused by problems with specific proteins in the body. Cystic fibrosis is one such disease, as are some neurodegenerative diseases such as Alzheimer's and Mad Cow. The study of proteins then, both their function and dysfunction, contributes to a critical knowledge base from which scientists may one day be able to "fix" problems in proteins- in other words, cure people of diseases caused by proteins. Researchers in the field of biophysics study the nature of the molecules, such as proteins, from which living organisms are composed.

The work of biophysics researchers is highly computationally intensive. Indeed, the computing demands of some biophysics projects are so great that the research may not be undertaken for lack of sufficient computing power available to the researcher (1). Biophysics researchers of The Scripps Research Institute (TSRI) have made significant progress in overcoming these limitations, thanks in part to the growth of grid technology and the foresight and pioneering work of computing scientists at the University of Virginia (UVa). Marty Humphrey, Assistant Professor in the Computer Science Department, is a long-standing leader of UVa's work in grid computing, and in bringing the benefits of

the grid to biophysics researchers on his campus.

Recent fruit of the work of Humphrey's team has been the installation of a Computational Biophysics portal on the UVa campus. This portal was created at UVa in conjunction with their membership in the National Science Foundation's (NSF's) National Partnership for Advanced Computational Infrastructure (NPACI). UVa's work with NPACI intersected with their participation in the National Science Foundation Middleware Initiative (NMI) Integration Testbed. Managed by the Southeastern Universities Research Association (SURA) on behalf of the NMI-EDIT Consortium², the testbed consisted of eight universities that participated in a closely coordinated effort to deploy and evaluate NMI technologies³. This article will discuss the synergist effects UVa's participation in these two programs has brought to their grid computing efforts and biophysics research.

UVa Background in Grid Computing

Computer scientists at UVa have significant experience in developing and working with grid technology. Marty Humphrey is a

² NSF Middleware Initiative-Enterprise and Desktop Integration Technologies (NMI-EDIT); <http://www.nmi-edit.org/>

³ <http://www1.sura.org/3000/NMI-Testbed.html>



member of the steering committee of the Global Grid Forum (GGF) and co-chair of the Security Area of the GGF. Humphrey and his team participated in the development of the grid operating system software Legion, principally developed by UVa colleague Prof. Andrew Grimshaw. Biophysicists at UVa worked with this team in 2000 to demonstrate Legion and highlight the significant benefits that grid technologies can have for biomolecular research. In their testing of Legion, a biomolecular researcher completed his computing work that would've taken a month to complete in less than *two days* by using grid technology (2).

In addition to Humphrey's work with Legion and NPACI, UVa's Jim Jokl is a leader in security and interoperability issues associated with the public key infrastructure (PKI). Their work in these areas provided UVa with the motivation to participate in the NMI Testbed. By participating in the NMI Testbed, UVa staff was able to more closely consider the integration of the NPACI Computational Biophysics portal with the campus infrastructure, particularly authentication infrastructures. In this sense, the NMI Testbed was a place for Humphrey and Jokl's to meld each of their areas of expertise in a way that not only has brought current benefits to their biophysics researchers, but is likely to contribute to how authentication on grids is done in the future⁴

(as we'll see in later in the "Lessons Learned" section).

The Study of Protein Folding

Amino acids are the building blocks of life, and specifically, of proteins. Our cells use amino acids to create proteins. Actually, it may be more accurate to say that it is thought that proteins *fold themselves* with the amino acids in cells. The process by which proteins are created is called "protein folding". The characteristics of the folds themselves depend on the cell environment that the protein is being created in (e.g., temperature, salinity). Through the folding process, proteins take on their unique three-dimensional shape that gives them their functionality. It is the errors that can occur in the folding process that cause disease such as Mad Cow (3, 4).

In order to prevent and correct protein folding errors, scientists must have a thorough understanding of the specifics of the folding process for each type of protein. Biomolecular researchers study the process of protein folding by modeling the folding process with simulation software packages (called "molecular dynamics codes"). CHARMM (Chemistry at Harvard Molecular Mechanics) and Amber (Assisted Model Building with Energy Refinement) are the primary simulation packages used by researchers (5). UVa researchers, for instance, have used CHARMM to study Protein L in its folded and unfolded states, to

⁴<http://www.cs.virginia.edu/~humphrey/papers/BridgeCAGridSecWorkshop2004.pdf>



look at its energy and entropy, and to generate a protein-folding landscape (6).

Easing the Study of Protein Folding at UVa

Computer scientists at UVa have worked to simplify the computing efforts of the biomolecular researchers on the UVa campus. Humphrey and his team have created and deployed a Computational Biophysics portal to assist their researchers working in biophysics. UVa staff implemented the Computational Biophysics portal as part of their work with NPACI. This portal makes the resources of the NPACI grid available to UVa researchers, and facilitates scientists' work in that it provides a simple user interface to access resources on the grid. By design, the grid is a network of disparate machines running different software in any number of geographic locations. While the base functionality of grid software provides the user with a layer of "insulation" from the complexity of this environment, portals provide an even more seamless, user-friendly computing interface from which they can run their applications across the grid. There are key technologies utilized for the portal and the NPACI grid from the GRIDS Center, including Condor-G⁵, Globus⁶, GridPort⁷, GSI OpenSSH⁸, MyProxy⁹, and MPICH-G2¹⁰. Collectively,

these NMI components are part of the NPACI distribution called the NPACiKage.

Lessons Learned

The use of the NMI/NPackage/NPACI Grid will greatly increase the number of users who might use biomolecular applications in a Grid environment. This fact, actually, provided UVa's staff with the motivation to participate in the NMI Testbed and specifically, to work with NMI components. UVa computing scientists were already deeply involved with grid computing software as the NMI Testbed "call for participation" went out. UVa researchers had been using grid software components in versions that were not the NMI compliant version. The NMI Testbed however, provided them the opportunity to work NMI compliant software into their on-going projects, such as protein folding research.

The decision to test NMI compliant versions and integrate them into their projects was an important one. UVa wanted to improve the quality of the NMI components in any way they could, because Humphrey believes that the NMI program is one of the best-run sources for middleware and UVa strongly wanted NMI to succeed. Testing NMI components in this context allowed UVa to find problems with the NMI software; the problems were then conveyed to the component authors and fixed. UVa also wanted to evaluate to what extent the NMI package could collectively provide the Grid middleware solution for applications such as CHARMM and Amber. UVa found that the

⁵ Condor-G information: <http://www.cs.wisc.edu/condor/>

⁶ Globus Toolkit information: <http://www-unix.globus.org/toolkit/>

⁷ GridPort information: <http://gridport.net/index.cgi>

⁸ GSI-OpenSSH information:

<http://grid.ncsa.uiuc.edu/ssh/>

⁹ MyProxy information:

<http://grid.ncsa.uiuc.edu/myproxy/>

¹⁰ MPICH-G2 information: <http://www3.niu.edu/mpi/>



pieces fit nicely together with minimal additions other than application-specific logic.

As suggested early in this article, in addition to enhancing the current research of UVa's biomolecular physicists, UVa's Humphrey and Jokl have another important lesson to share with others. Humphrey states that their work in the NMI Testbed has made it clear that to him that when folks set about deploying NMI on their campuses, that they should attempt as much as possible to reuse existing authentication infrastructures in their grid deployments.

Many campuses, businesses, and science-based organizations are building their own grids. While the function of a grid is to provide a single sign-on, that sign-on allows access only to the resources of that specific grid. As NMI becomes more mature and stable, even more grids will be built. The default temptation is for each "application grid" to set up its own authentication infrastructure (i.e., CA or certificate authority) and manage the id's and passwords of those that wish to use the grid. However, many researchers and others will want to access more than a single grid. If each grid has implemented its own authentication, inevitably each user will end up hand-managing multiple credentials (sign-ons), which is tedious and ultimately insecure.

As part of their participation in the NMI Testbed, UVa's Jokl and Humphrey have

undertaken the first steps in re-using their campus authentication infrastructures across grids. Jokl and Humphrey are working with NPACI to have NPACI's policies changed so that they allow a local campus's credentials (either from a Campus CA or a NMI-EDIT [Shibboleth](#)^{®11} server on campus) to be used to access the NPACI grid. The NMI Testbed project period ran out before these NPACI's policies could be changed. However Marty Humphrey and Jim Jokl, as part of a subcontract they have to the San Diego Supercomputing Center (SDSC), plan to continue to work on this issue into the fiscal year 2005. With the expected growth of grids, indeed, with the certainty of the continued expansion of grid computing technology across the Internet, simplifying grid authentication infrastructures will have a huge impact for those that deploy grids and for grid users.

More Information

For more information about UVa Protein Folding on the Grid, contact Marty Humphrey at humphrey@cs.virginia.edu.

References

- (1), (2), (6) "Studying Protein Folding on the Grid: Experiences using CHARMM on NPACI Resources under Legion", Natrajan, A, Crowley, M, Wilkins-Diehr, N, Humphrey, MA, Fox, AD, Grimshaw, AS, Brooks III, CL.
- (3) http://en.wikipedia.org/wiki/Protein_folding
- (4) <http://folding.stanford.edu/>
- (5) http://npacigrid.npaci.edu/case_studeis_protein_folding.html

¹¹ Shibboleth information: <http://shibboleth.internet2.edu>



Links of Interest

The University of Virginia <http://www.virginia.edu/>

Amber <http://www.uscf.edu>

CHARMM <http://yuri.harvard.edu/>

Global Grid Forum (GGF) <http://www.gridforum.org/>

GRIDS Center <http://www.grids-center.org/>

Humphrey, Marty <http://www.cs.virginia.edu/~humphrey/>

Legion <http://legion.virginia.edu/>

National Partnership for Advanced Computational Infrastructure (NPACI) <http://www.npaci.edu/>

NMI-EDIT <http://www.nmi-edit.org/>

NMI Integration Testbed Program <http://www1.sura.org/3000/NMI-Testbed.html>

NSF Middleware Initiative <http://www.nsf-middleware.org/>

The Scripps Research Institute (TSRI) http://www.scripps.edu/e_index.html

“Studying Protein Folding on the Grid: Experiences using CHARMM on NPACI Resources under Legion” <http://legion.virginia.edu/papers/hpdc01.pdf>

UVa Computational Biophysics portal
<https://wumpus.cs.virginia.edu/NPACIComputationalBiophysicsPortal/>