Grid-enabled High-throughput in silico Screening against Influenza A Neuraminidase

H.-C. Lee, J. Salzemann, N. Jacq, H.-Y. Chen, Li-Yung Ho, I. Merelli, L. Milanesi, Vincent Breton, S. C. Lin, Y.-T. Wu

Grid-enabled High-throughput *in silico* Screening against Influenza A Neuraminidase

Hurng-Chun Lee, Jean Salzemann, Nicolas Jacq, Hsin-Yen Chen, Li-Yung Ho, Ivan Merelli, Luciano Milanesi, Vincent Breton, Simon C. Lin, Ying-Ta Wu

***Abstract*—Encouraged by the success of the first EGEE biomedical data challenge against malaria (WISDOM) [1], the second data challenge battling avian flu kicked off in April 2006 to identify new drugs for the potential variants of the Influenza A virus. Mobilizing thousands of CPUs on the Grid, the 6-weeks long high-throughput screening activity has fulfilled over 100 CPU years of computing power and produced around 600 Gigabytes of results on the Grid for further biological analysis and testing. In the paper, we demonstrate the impact of a world-wide Grid infrastructure to efficiently deploy large scale virtual screening [2] to speed up the drug design process. Lessons learned through the data challenge activity are also discussed.**

Index Terms—data challenge, EGEE, BioinfoGRID, grid infrastructure, virtual screening, drug discovery, avian flu, neuraminidase.

I. Introduction

The potential for re-emergence of influenza pandemics has been a great threat since the report that the avian influenza A virus (H5N1) could acquire the ability to be transmitted to humans. Indeed, an increase of transmission incidents suggests a risk of human-to-human transmission [3]. Furthermore, the report of development of drug resistance variants [4] is another potential concern. Two of the present drugs (oseltamivir and zanamivir) were discovered through structure-based drug design targeting influenza neuraminidase (NA), a viral enzyme that cleaves terminal sialic acid residue from glycoconjugates of cell surface. The action of NA is essential for virus proliferation and infectivity; therefore, blocking its activity generates antivirus effects. To date, there is no NA subtype one (N1) available for structural study. To minimize non-productive trial-and-error approaches and to accelerate the discovery of novel potent inhibitors, medical chemists take advantage of modeled NA variant structures and structure-based design.

A key work in structure-based design is to model complexes of candidate compounds to structures of receptor binding sites. The computational tools for the work are based on molecular docking engines, such as AutoDock [5], to carry out a quick conformation search of small compounds in the binding sites, fast calculation of binding energies of possible binding poses, prompt selection for the probable binding modes, and precise ranking and filtering for good binders. Although docking engines can be run automatically, one needs to control the dynamic conformation of the macromolecular binding site (rigid or flexible) and the spectrum of the screening small organics. Such consideration will decide the complexity of the modeling system. This process is characterized by computational and storage loads which pose a great challenge to resources that a single institute can afford.

In April and May 2006, the second biomedical data challenge of the EGEE project led by Academia Sinica in Taiwan, CNRS-IN2P3 in France and the European SSA BioinfoGRID project coordinated by CNR-ITB in Italy was kicked off to tackle the computational challenge of screening about 300,000 compounds selected from ZINC [6] and a chemical combinatorial library against 8 variants of neuraminidases predicted by homology method. Using AutoDock as the docking engine, the computation requires over 100 years when run on an average PC. In order to compress the overhead so that biomedical chemists can have best response to instant threads while the mutation of the virus happens, more than 2000 CPUs in the EGEE Grid infrastructure have been mobilized to perform large scale distributed virtual screening during 6 weeks. About 600 Gigabytes of output data have been produced and archived on the Grid with one additional backup.

Beside the biological goal of reducing the time and cost of the initial investment on structure-based drug design, there are

Manuscript received July 10, 2006. This work was supported in part by AuverGrid, TWGrid and EGEE projects. EGEE is a project funded by the European Union under contract INFSO-RI-508833. The TWGrid is funded by the National Science Council (NSC), Taiwan. Auvergrid is a project funded by the Conseil Regional d'Auvergne.

H. C. Lee, L.-Y. Ho, H.-Y. Chen, S. C. Lin and Y. T. Wu are with Academia Sinica, No. 128, Sec. 2, Academic Rd., NanKang, Taipei 115, Taiwan. E-mail: {hclee, liyungho, hychen, sclin, ywu}@gate.sinica.edu.tw.

J. Salzemann, N. Jacq and V. Breton are with CNRS IN2P3, Laboratoire de Physique Corpusculaire, Campus des Cézeaux, 24 av. des Landais, 63177 Aubière, France, {salzeman, jacq, breton}@clermont.in2p3.fr

I.Merelli and L. Milanesi are with CNR-ITB, CNR-Institute for Biomedical Technologies, Via Fratelli Cervi 93, 20090 Segrate (Milan), Italy, {merelli, milanesi}@itb.cnr.it

two Grid technology objectives for this activity: one is to improve the performance of the *in silico* high-throughput screening (HTS) environment based on what has been learnt in the previous challenge against Malaria (WISDOM) [7]; the other is to test another environment which enables users to have efficient and interactive control of the massive molecular dockings on the Grid. Therefore, two Grid tools were used in parallel in the second data challenge. An enhanced version of WISDOM high-throughput workflow was designed to achieve the first goal and a light-weight framework called DIANE [8] was introduced to carry a significant fraction of the deployment for implementing and testing the new scenario.

The paper is organized as follows. The second section briefly introduces the Grid environments on which the data challenge was executed. In section 3, the two Grid tools used to generate the data challenge are presented. In section 4, the data challenge activity is described, particularly its preparation, deployment and execution. In section 5, the discussion focuses on the general statistics, efficiency and issues we observed and experienced in the data challenge. The last section draws the final conclusions.

I. The Grid Infrastructure

Three infrastructures were used to achieve the deployment: AuverGrid [9], TWGrid [10] and EGEE [11]. In this section, we are describing them briefly.

AuverGrid is regional grid deployed in the French region Auvergne. Its goal is to explore how a grid can provide the resources needed for public and private research at a regional level. With more than 800 CPUs available at 12 sites, AuverGrid hosts a variety of scientific applications from particle physics to life science, environment and chemistry.

TWGrid is responsible for operating a Grid Operation Center in Asia-Pacific region. Apart from supporting the world-wide Grid collaboration in high-energy physics, TWGrid is also in charge of federating and coordinating regional Grid resources to promote Grid technology to the e-Science activities (e.g. life science, atmospheric science, digital archive, etc.) in Asia.

The Enabling Grids for E-sciencE project (EGEE) brings scientists and engineers together from more than 90 institutions in over 30 countries world-wide to provide a seamless Grid infrastructure for e-Science that is available for scientists 24 hours-a-day. The EGEE Grid consists of over 30,000 CPU available to users 24 hours a day, 7 days a week. 5 Petabytes of storage are available, and on average 20,000 concurrent jobs are executed. Expanding from originally two scientific fields, high energy physics and life science, EGEE now integrates applications from many other scientific fields, ranging from geology to computational chemistry.

To efficiently operate the distributed resources as a whole system, the EGEE Grid middleware [12] provides a User Interface (UI), a Workload Management System (WMS) relying on resource broker machines, a Data Management System (DMS), an Information System (IS), and several monitoring and application deployment tools based on the Grid Security Infrastructure (GSI). All the Grid activities and resource sharing within EGEE are operated and coordinated within the scope of Virtual Organizations (VOs) [13], virtual communities across laboratories and institutes around the world.

The data challenge against avian flu was officially supported by the biomedical VO of the EGEE and BioinfoGRID projects. Resources from AuverGrid and TWGrid were explicitly allocated to complement the EGEE resources.

II. The Grid Tools

A. The WISDOM production environment

A large scale deployment requires the development of an environment for job submission and output data collection. A number of issues need to be addressed to achieve significant acceleration from the grid deployment:

- The amount of data moved around at job submission has an impact on Grid performances. As a consequence, the files providing the 3D structure of targets and compounds should preferably be stored on grid storage elements in preparation for the data challenge.
- The rate at which jobs are submitted to the grid resource brokers must be carefully monitored in order to avoid their overload. The job submission scheme must take into account this present limitation of the EGEE brokering system.
- The Grid submission process introduces significant delays for instance at the level of resource brokering. The jobs submitted to the grid computing nodes must be sufficiently long in order to reduce the impact of this middleware overhead.

The WISDOM production environment was designed to achieve production of a large amount of data in a limited time using EGEE, AuverGrid and TWGrid middleware services. Three packages were developed in Perl and Java. Their entry points are a simple command line tool. The first package installs the application components (software, compounds database…) on the grid computing nodes. The second package tests these components. The third package monitors the submission and execution of the WISDOM jobs.

The environment was improved to address limitations and bottlenecks identified during the first data challenge against malaria deployed in the summer of 2005 on the EGEE infrastructure. For instance the number of resource broker machines and the rate at which the jobs were submitted on them were extended to avoid their overloading. Another improvement concerned the resubmission process after a job failure which was redesigned to avoid a “sink-hole” effect on a failing grid computing node. Automatic resubmission was replaced by the manual intervention of the WISDOM production user.

B. The DIANE framework

DIANE is a lightweight distributed framework for parallel scientific applications in a master-worker model. It assumes

that a job may be split into a number of independent tasks which is a typical case in many scientific applications. It has been successfully applied in a number of applications ranging from image rendering to data analysis in high-energy physics.

As opposed to standard message passing libraries such as MPI [14], the DIANE framework takes care of all synchronization, communication and workflow management details on behalf of the application. The execution of a job is fully controlled by the framework which decides when and where the tasks are executed. Thus the application is very simple to program and contains only the essential code directly related to the application itself without the need for networking details.

Aiming to efficiently bridge underlying distributed computing environments and application centric user interface as illustrated in Fig 1, DIANE itself is a thin software layer which can easily work on top of more fundamental middleware such as LSF, PBS or the Grid Resource Brokers. It may also work in a standalone mode and does not require any complex underlying software.

As a framework, DIANE provides an adapter for applications. Fig 2 shows the template of DIANE application plug-ins. A complete DIANE application plug-in should implement three major Python objects: the *Planner* and the *Integrator* objects implement the job splitting and result merging, respectively; while the logic of the *Worker* object concentrates on the execution of the individual task. When a DIANE job is started by a user, both the *Planner* and the *Integrator* objects are invoked by a master agent usually executed on the user's desktop, and typically the worker agents are submitted to run on distributed CPUs such as the Grid worker nodes.

Once the worker agent is launched, first it registers itself with the master agent. In the second step, a channel is established for pulling the tasks from the queue held by the master agent. When the individual task is done by the worker agent, the result is returned and merged on the master. The pulling-executing- returning cycle will iterate until all the tasks are accomplished. The same channel is also used to profile the worker agent's health and to support user interaction with the task. The whole DIANE framework is written in Python and the communication between the master agent and the worker agents is based on the CORBA protocol [15].

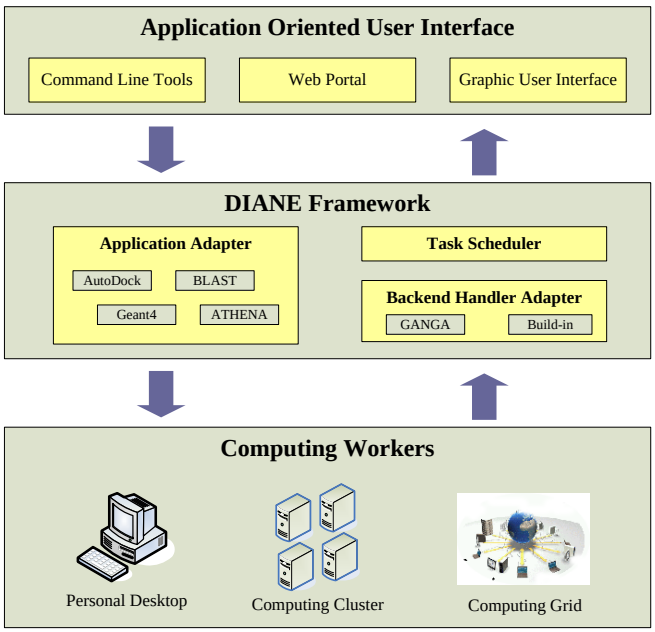

Fig 1 The DIANE framework sitting on top of a variety of computing environments provides a fast integration of distributed and heterogeneous computing resources. It hides the scheduling details of application distribution so that on top of it, application oriented user interfaces could be easily developed.

Since the DIANE framework takes care of the control of the communication and the workflow on behalf of the application, implementing an AutoDock adaptor for DIANE costs approximately 3 days and the effort is less than 500 lines of Python codes.

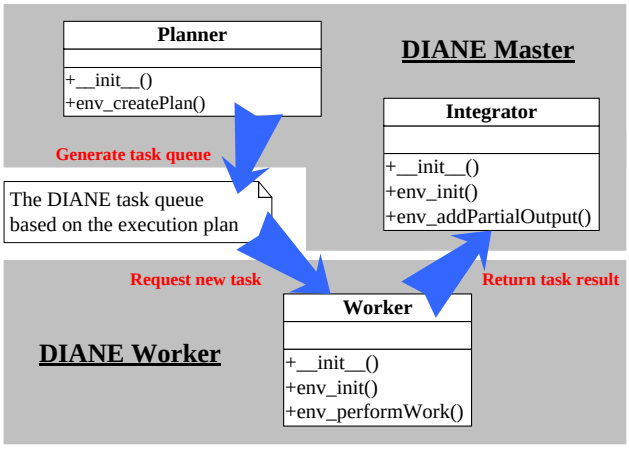

Fig 2 the template of DIANE application plug-ins as well as the cooperation model between the three major objects: Planner, Worker and Integrator.

III. The Data Challenge

The name of *Data Challenge* is inspired from the large-scale exercise on the World-wide LHC Computing Grid (WLCG) which aims at processing a huge number of collision events produced by the Large Hardron Collider (LHC). Instead of processing the physics events, the biomedical data challenge deals with the biomedical data, for example medical image processing or virtual screening. The avian flu data challenge is

the second official biomedical data challenge of the EGEE project. The previous data challenge against malaria [7] was done during the summer of 2005 and saw over 46 million compounds docked in 6 weeks.

Input for the avian flu data challenge consists of 8 protein targets predicted from the neuraminidases subtype 1 (N1) to simulate the possible mutations of the H5N1 virus and 308,585 chemical compounds selected from ZINC and a chemical combinatorial library. By dividing the 308,585 chemical compounds into 2 subsets, the whole data challenge activity was broken down to 16 instances; each instance corresponded to the dockings of an N1 variant against the compounds in one of the 2 subsets. To avoid that concurrent executions of all the instances overload the Grid system and reduce the Grid efficiency, the initialization time of each instance was well scheduled.

The majority of the data challenge instances were executed using the WISDOM production environment since its scalability had been demonstrated already in the first data challenge. Due to the fact that the CPU wall time of most of the Grid computing elements are restricted to 24 hours, the Grid jobs submitted by WISDOM were carefully partitioned to prevent from running over this limitation. Taking into account the approximation that the computing time of each single docking is about 30 minutes[2], each WISDOM job was prepared to run on 40 dockings. Thus each instance represented 7715 Grid jobs. In order to balance the load on the Grid Workload Management System, WISDOM submitted the jobs to 18 resource brokers in a round-robin order.

In parallel with the WISDOM activity, DIANE was used to run as many dockings as it could handle during the data challenge activity. To avoid the resource competition with WISDOM, DIANE took only a small fraction of the available resources. Unlike WISDOM, how the job is split into independent DIANE tasks plays and important role in the overall distribution efficiency of a DIANE job. As the estimated elapsed time of each docking is significantly longer than the startup overhead of the task, each DIANE task was defined to corresponding to the docking of one compound. As a master-worker model, DIANE submitted worker agents instead of docking tasks to the Grid. As a consequence, the wall time limitation affects the lifetime of the worker agents and more worker agents need to be submitted once the limitation is reached. During the data challenge, a DIANE master was maintained on the UI to hold a queue of the waiting docking jobs and a separate process for submitting DIANE worker agents was manually triggered. This strategy allowed using more CPU power to ramp up the docking throughput without interfering with the running master. The result of each docking was interactively returned back to the Grid UI once the task was successfully completed. All the results were also concatenated and archived into the Grid.

To share the data challenge results for further biological analysis,, about 120,000 files in total were archived in Taiwan and in France. The centralized LCG File Catalog (LFC) system was used to index all the files distributed on the Grid.

Before data challenge kick-off, the compounds were pre-staged on 3 Grid SEs, and the Autodock executable was widely deployed on most of the available Grid CEs. Based on what has been learnt in the previous data challenge, the deployment work including the prediction of the N1 variants took about 1 month.

IV. Discussion

A. General Statistics

Table 1 and Table 2 summarize the data challenge deployments using WISDOM and DIANE environments, respectively.

TABLE 1
STATISTICAL SUMMARY OF THE WISDOM ACTIVITY

Total number of completed dockings	$2 * 10^6$
Estimated duration on 1 CPU	88.3 years
Duration of the experience	6 weeks
Cumulative number of Grid jobs	54,000
Maximum number of concurrent CPUs	2,000
Number of used Computing Elements	60
Crunching factor	912
Approximated distribution efficiency	46%

TABLE 2
STATISTICAL SUMMARY OF THE DIANE ACTIVITY

Total number of completed dockings	308,585
Estimated duration on 1 CPU	16.7 years
Duration of the experience	4 weeks
Cumulative number of Grid jobs	2585
Maximum number of concurrent CPUs	240
Number of used Computing Elements	36
Crunching factor	203
Approximated distribution efficiency	84%

During the data challenge, the WISDOM activity has distributed 54,000 jobs on 60 Grid CEs. The 6-weeks activity has covered the computing power of about 88 CPU years and has docked about 2 million pairs of target and chemical compounds. Due to the fact that the Grid resources were used by other VOs during the data challenge, a maximum of 2000 CPUs were concurrently running at the same time. For the DIANE part, we were able to complete 308,585 docking runs (i.e. 1/8 of the whole challenge) in 30 days using the computing resources of 36 Grid CEs. A total number of 2580 DIANE worker agents have been running as Grid jobs during that period and 240 of them were concurrently maintained by the

[2] The measurement was done on a PC with one Xeon 2.8 GHz CPU and 2 Gigabytes physical memory.

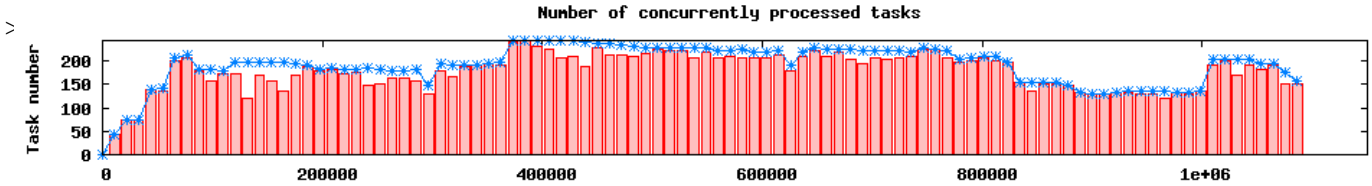

Fig 4 the resource utilization of a DIANE job. The solid curve with crosses illustrates the number of CPUs available for doing the dockings; while the bars indicate the concurrent executing dockings (i.e. the utilized CPUs).

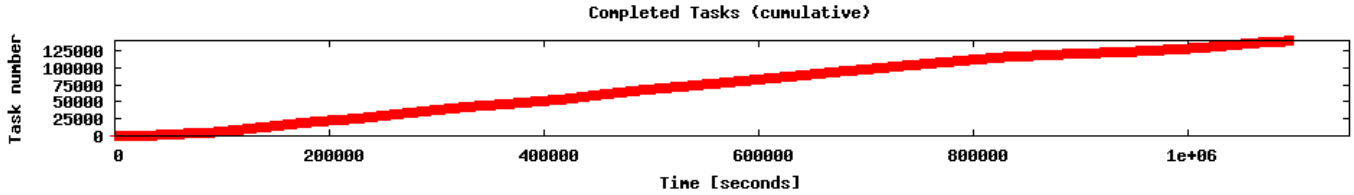

Fig 3 the docking throughput of a DIANE job. The curve shows the cumulative number of the completed dockings during the job lifetime of about 2 weeks.

DIANE master. The distribution of those Grid jobs in terms of the regions of the world is shown on Fig 4. About 600 Gigabytes of date have been produced on the Grid during the data challenge.

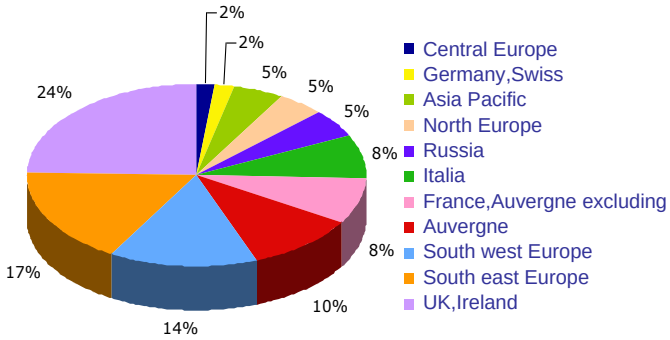

Fig 4 the distribution of the Grid jobs in different region.

B. *Efficiency and throughput*

Since Grid is a dynamic system in which the status of resources is changed without central control, transient problems occur which cause job failures. In the WISDOM activity, about 83% of the jobs were reported as successfully finished according to the status logged in the Grid Logging and Bookkeeping system (LB); the observed failures were mainly due to errors at job scheduling time because of mis-configuration of Grid Computing Elements (CE). However, the success rate went down to 70% after checking the content of the data output file. The main cause for these failures was frequent last-minute error in the transfer of results to the Grid Storage Elements. Compared to the previous data challenge, improvement is significant as the observed success rates were respectively 77 and 63%. The last-minute error in output data transfer is particularly expansive since the results are no longer available on the Grid Worker Node (WN) although they might have been successfully produced.

In DIANE, a similar job failure rate was also observed; nevertheless, the failure recovery mechanism in DIANE automated the re-submission and guaranteed a fully complete job. On the other hand, the feature of interactively returning part of the computing efforts during the runtime (e.g. the output of each docking) also introduces a more economical way of using the Grid resources.

For the instances submitted using the WISDOM production environment, the overall crunching factor was about 912.

The corresponding distribution efficiency defined as the ratio between the overall crunching factor and the maximum number of concurrently running CPUs, was estimated to 46%. This is due to the known issue of long job waiting time in the current EGEE production system.

The task pull model adopted by DIANE allows the isolating of the scheduling overhead of the Grid jobs and is therefore expected to achieve a better distribution efficiency. During the data challenge, DIANE was able to push the efficiency to higher than 80%. Fig 4 presents the resource utilization of a DIANE job. Although DIANE was not tested in a very large scale like WISDOM, the good resource utilization shown in **Erreur ! Source du renvoi introuvable.** still provides details on the improvement. The cumulative plot of the completed dockings in **Erreur ! Source du renvoi introuvable.Erreur ! Source du renvoi introuvable.** also demonstrated that a constant throughput can be effortlessly maintained for few weeks using the task pull model.

Because of the highly scalable nature of the WISDOM framework, high throughput docking could be achieved at a rate of 2 seconds per docking. As DIANE was handling not more than a few hundred concurrent jobs, its throughput was limited to about one docking every 10 seconds.

A. *Standing issues*

1) *Issues related to the Grid middleware*

The scheduling costs introduced by the current middleware on the Grid jobs are significantly high. One of the reasons is that a sequential and continuous job submission to the Grid will heavily load the Workload Management System (WMS), therefore the WMS will then take more time for resource match-making and job dispatching to the Grid Computing Elements (CE). Another problem comes from the fact that WMS is not aware of the resource usage priority given by the resource sharing policy implemented in the local queuing

system on the CE. Without having the information published by the CE, the Information System might wrongly guide the WMS to send jobs to a CE on which the owner of the jobs has relatively low priority on the usage of resources. Relevant activities concerning these Grid scheduling issues have been held in the preparation of the next generation EGEE middleware.

Mission critical applications (e.g. disease diagnosis, drug discovery, etc.) running on the Grid require different levels of Quality of Service (QoS). Taking the example of the data challenge, the throughput is one of the key QoS parameters as time may become a critical factor to address emerging diseases. The avian flu virus might spread-out at an un-expected speed once the variant with the ability of human-to-human transmission comes out. According to the definition of the QoS taxonomy [16], the QoS in the current EGEE middleware is implemented in a soft way based only on the ranking and match-making mechanism provided by the WMS system. The WMS relying highly on the IS has no way to guarantee that its resource selection will meet user QoS requirement. Thus how to ensure the QoS within the current Grid middleware is still an open question. In some sense, site functional tests [17] and some utilities made available at User Interface provide some ad-hoc solutions for users to check the status of the Grid sites before job submission; however, a more promising solution so far has been to adopt the negotiation protocols for service level agreement.

To manage a large scale production, the Grid monitoring and accounting tool is very helpful for tracing the progress as well as the failures of the jobs. Several tools [18] are delivered as part of the EGEE middleware for monitoring the Grid activities in different aspects. Out of those tools, the GridICE [19] and the GOC accounting system [20] provide the statistical data in the views of VOs, which is more intuitive than the resource-centric information given by the other tools in monitoring the data challenge progress. However, the sensors producing job-monitoring information deployed on every site are not correctly configured everywhere. This yields partial information and makes the report difficult to interpret.

2) Issues related to the WISDOM production environment

The WISDOM production environment achieved large scale deployment. But the failure rate is still high despite environment improvements to address issues identified during the previous data challenge. The main remaining limitations are related to the performances of the resource broker machines and the grid computing nodes stability.

Automatic resubmission applied during the first data challenge was a cause of failures and consequently a time-consuming correction task for the job supervisor. Resubmission by hand allows the process to be checked precisely but limits the building of an automatic pipeline of grid-enabled virtual screening. An issue for a next data challenge is to improve the WISDOM production environment to manage efficient automatic resubmission with only relevant resource brokers and grid computing nodes.

An idea could be to develop a learning module to register failed and efficient resource brokers and nodes. The module could make the information available during the submission and the monitoring process of the WISDOM production environment to modify the requirements of new submitted and resubmitted jobs.

3) Issues related to the DIANE platform

The scalability issue of the DIANE framework is due to the fact that the DIANE master needs to keep the connections with the distributed DIANE workers for task dispatching and worker health checking. Performance evaluation during the data challenge showed that the current implementation of the DIANE master is restricted to handle few hundred DIANE workers at the same time. The main reason for this restriction is still under investigation. An alternative solution is to adopt stateless protocols (e.g. web services [21] or stateless CORBA); however, this will introduce an overhead for establishing every connections and a tradeoff should be made between the performance and scalability.

For instance, to give users more flexible control over their DIANE jobs, the master of DIANE is usually executed on the Grid User Interface. This feature will turn into a performance issue while the payload of result integration is high. The heavily loaded integration process will affect the performance of the UI. A possible approach to address this issue is to run the DIANE master as a Grid job on a Grid Worker Node; however, one should make sure that the master is always started before the workers and the network connectivity between two Grid WNs becomes yet another problem.

V. Conclusion

We have performed a large-scale high-throughput *in silico* screening on the Grid in search for potential drugs against the predicted variants of the avian flu virus, H5N1. Using three Grid infrastructures (AuverGrid, EGEE, TWGrid), we have successfully reduced the duration of the virtual screening process from over 100 years to 6 weeks. High throughput *in silico* docking was achieved with up to one docking every 2 seconds. The results are now under analysis and the outcome will help biomedical chemists to reduce the cost of the first investment in the process of structure-based drug design.

Two different Grid tools were used to execute the data challenge. The WISDOM and DIANE production environments have been described and compared. During this second data challenge, we proved again that the WISDOM production environment is capable of controlling a high-throughput screening with a reduced preparation effort. We demonstrated that the DIANE light-weight framework offered an improved distribution efficiency as well as a steady throughput of the distributed molecular dockings on the Grid.

Several issues related the Grid middleware as well as the two Grid production environments have been highlighted. Investigations and discussions with the developers are taking place in the preparation of the next data challenge against neglected diseases that will take place in the fall of 2006 in the framework of the EGEE and BioinfoGRID [22] projects.

ACKNOWLEDGMENTS

The authors express particular thanks to the site managers in EGEE, TWGrid, AuverGrid, BioinfoGRID for operational support, the LCG ARDA group for the technical support of DIANE, and the Biomedical Task Force and Embrace for its participation to the WISDOM deployment. The following institutes contributed computing resources to the data challenge: ASGC (Taiwan); NGO (Singapore); IPP-BAS, IMBM-BAS and IPP-ISTF (Bulgaria); CYFRONET (Poland); ICI (Romania); CEA-DAPNIA, CGG, IN2P3-CC, IN2P3-LAL, IN2P3-LAPP and IN2P3-LPC (France); SCAI (Germany); CNR-ITB and INFN (Italy); NIKHEF, SARA and Virtual Laboratory for e-Science (Netherlands); IMPB RAS (Russia); UCY (Cyprus); AUTH FORTH-ICS and HELLASGRID (Greece); RBI (Croatia); TAU (Israel); CESGA, CIEMAT, CNB-UAM, IFCA, INTA, PIC and UPV-GryCAP (Spain); BHAM, University of Bristol, IC, Lancaster University, MANHEP, University of Oxford, RAL and University of Glasgow (United Kingdom).

REFERENCES

[1] WISDOM: Wide In Silico Docking On Malaria, http://wisdom.eu-egee.fr
[2] W. P. Walters, M. T. Stahl, and M. A. Murcko, "Virtual Screening - an Overview", *Drug Discovery Today*, 3:160-178, 1998
[3] K. S. Li, Y. Guan, J. Wang, G. J. D. Smith, K. M. Xu, L. Duan, A. P. Rahardjo, P. Puthavathana, C. Buranathai, T. D. Nguyen, A. T. S. Estoepangestie, A. Chaisingh, P. Auewarakul, H. T. Long, N. T. H. Hanh, R. J. Webby, L. L. M. Poon, H. Chen, K. F. Shortridge, K. Y. Yuen, R. G. Webster and J. S. M. Peiris, "Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia", *Nature* 430:209-213, 2004
[4] M. D. de Jong, T. T. Tran, H. K. Truong, M. H. Vo, G. J. Smith, V. C. Nguyen, V. C. Bach, T. Q. Phan, Q. H. Do, Y. Guan, J. S. Peiris, T. H. Tran and J. Farrar, "Oseltamivir Resistance during Treatment of Influenza A (H5N1) Infection", *N. Engl. J. Med.*, 353(25):2667-72, 2005.
[5] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, "Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function", *J. Computational Chemistry*, 19:1639-1662, 1998.
[6] Irwin and Shoichet, *J. Chem. Inf. Model.*, 45(1):177-82, 2005
[7] V. Breton, N. Jacq, and M. Hofmann, "Grid added value to address malaria", *Proceedings of the 6-th IEEE/ACM CCGrid conference*, 2006
[8] DIANE: Distributed Analysis Environment, http://cern.ch/diane
[9] AuverGrid, http://www.auvergrid.fr
[10] TWGrid, http://www.twgrid.org
[11] EGEE: Enabling Grids for E-science in Europe, http://public.eu-egee.org
[12] LCG-2 Middleware Overview, https://edms.cern.ch/file/498079/0.1/LCG-mw.pdf
[13] I. Foster, C. Kesselman and S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations", *Int. J. Supercomputer Applications*, 15(3), 2001
[14] W. Gropp and E. Lusk, "Dynamic process management in an MPI setting", *Proceedings of the 7th IEEE Symposium on Parallel and Distributed Processing*, October, 1995
[15] OMG, http://www.omg.org/gettingstarted/corbafaq.htm
[16] K. Krauter, R. Buyya and M. Maheswaran, "A Taxonomy and Survey of Grid Resource Management Systems for Distributed Computing", *Softw. Pract. Exper.*, 32:135-164, 2002
[17] The Site Functional Test (SFT) of EGEE/LCG production environment, https://lcg-sft.cern.ch/sft/lastreport.cgi
[18] GOC Grid Monitoring, http://goc.grid-support.ac.uk/gridsite/monitoring/
[19] S. Andreozzi, N. De Bortoli, S. Fantinel, A. Ghiselli, G.L. Rubini, G. Tortone and M.C. Vistoli, "GridICE: a Monitoring Service for Grid Systems", *Future Generation Computing Systems*, Elsevier, 21(4):559-571, 2005
[20] GOC Accounting Services, http://goc.grid-support.ac.uk/gridsite/accounting/
[21] G. Alonso, F. Casati, H. Kuno, and V. Machiraju, *Web Services*, Springer Verlag, 2003
[22] http://www.bioinfogrid.eu/

Hurng-Chun Lee received his Master's degree in Physics from Chung-Yuan Christian University, Chung-Li, Taiwan in 1999. He was working on computational physics and bioinformatics in Academia Sinica Computing Centre (ASCC) from 1999 to 2003. During that period, he developed a parallel program for gene annotation implementing the CRASA algorithm, and a web-based portal environment to integrate the distributed bioinformatics computing resources supported by the National Resource Project for Genomic Medicine (NRPGM). He is working on the developments of the DIANE and GANGA frameworks with LCG-ARDA group at CERN. He is also a project manager of the Grid Computing Team of Academia Sinica in Taiwan and responsible for the integration and the deployment of the Grid applications.

Jean Salzemann graduated in 2002 with an IT Engineering degree from Ecole d'Ingénieur en Informatique pour l'Industrie at Tours. He started to work as a developer building management software. He joined Vincent Breton's team in 2004 at the French National Centre for Scientific Research (CNRS), where he started to work on grid environments especially managing grid middleware deployments. During 2005 he worked in the French project RUGBI, developing grid services and components. He is now member of the EMBRACE European project, making technology recommendations in the WP3 work package.

Nicolas Jacq is currently a PhD student for the French National Centre for Scientific Research (CNRS) and the IT society Communication & Systèmes at the Laboratoire de Physique Corpusculaire of Clermont-Ferrand, France. In 2000, he completed his biological engineering degree and worked for 3 years on the DataGrid project at the Laboratoire de Biologie des Protistes in Clermont-Ferrand, France. His project is the development of bioinformatics services in a grid environment. His main use case is the deployment of a virtual screening platform at a large scale on neglected and emerging diseases in the EGEE project.

Li-Yung Ho received Bachelor degree of Mathematics and Master's degree of Physics from National Chung-Cheng University in 2000 and 2002 respectively. He was employed by Academia Sinica Computing Center for a bioinformatics project from 2003 to 2005 and joined the Grid Computing Team of Academia Sinica in 2006. He is in charge of deploying biomedical applications on the Grid.

Hsin-Yen Chen has been working on the IT service of scientific computing at Academia Sinica since 1991. He was in charge of coordinating the high

performance and bioinformatics computing. He coordinated a bioinformatics IT project to develop a portal-based high-throughput computing environment for the NRPGM project from 2000 to 2004. He is now responsible for the coordination of the deployment of HEP and Biomedical applications in the WLCG/EGEE project. He is also interested in the research of the catalyst on those metal oxide surfaces with the density functional theory calculation.

Ivan Merelli received is MSc degree in Biomedical Engineering from the Polytechnic University, Milan, Italy, in 2003 with a thesis about molecular surface modeling and analysis. His research activities include the development of software for sequence based Genomics and for structural Proteomics research, with particular interest in protein-protein interaction. He works actively on the high performance implementation of Bioinformatics components using parallel programming and distributed platforms. Currently he works at the Institute of Biomedical Technology of the National Research Council, Italy, for the European "Specific Support Action for Bioinformatics in EGEE - BioinfoGRID" and the FIRB MIUR project "Italian Laboratory for Bioinformatics Technologies - LITBIO".

Luciano Milanesi is currently researcher of the Italian National Research Council – Institute of Biomedical Technologies (CNR-ITB). He became head of the Bioinformatics and Molecular Modelling Division of the Institute of Biomedical Technologies CNR in 1988. He has been teaching Informatics and Bioinformatics courses at Milan University since 2001. His main research activities include the Human Genome Project, developing tools for the genome sequence analysis and prediction of gene structure in different organisms, promoter prediction, gene expression analysis and the development of databases and data mining. He is group leader for the Bioinformatics at CISI "Centre for Bio-molecular Interdisciplinary Studies and Industrial applications". He has been the principle investigator for the European Project: TRADAT "TRAnscription Database and Analysis Tools", ORIEL "an Online Research Information Environment for the Life Sciences" and he is the coordinator of the European BIOINFOGRID project: "Bioinformatics Grid Applications for life science", the coordinator of the Italian LITBIO project: "Laboratory of Bioinformatics Technologies", and CNR representative in the EGEE II European Project. He is Editorial Board Member of Briefings in Bioinformatics and the IEEE Transaction in NanoBiosciences journals. He is the author of more than 140 publications in the field of Bioinformatics, Systems Biology and Medical Informatics.

Vincent Breton received his Engineer degree from Ecole Centrale de Paris in 1985 and his PhD in Nuclear Physics from the University of Paris XI- Orsay in 1990. From 1990, he has been a research associate at the French National Centre for Scientific Research (CNRS). In 2001, he founded a research group (http://clrpcsv.in2p3.fr) on the application to biomedical sciences of the IT technologies and tools used in high energy physics. Co-founder of the GATE collaboration (http://opengate.in2p3.fr) gathering more than 20 research laboratories around the world, co-founder of the Healthgrid and WISDOM initiatives, chairman of the first European conferences on grids for health in January 2003 and January 2004, he is involved in several FP6 European projects dealing with grids for life sciences and healthcare (Embrace, EGEE-II, BioinfoGRID, Share).

Simon C. Lin is in charge of the Academia Sinica Grid Computing Centre (ASGC) and acting as the committee member of Overview Board, Management Board and Grid Deployment Board of the LHC Computing Grid (LCG) project led by CERN. He is also responsible for the Asia Federation and a member of PMB in Enabling Grid for E-sciencE (EGEE) project. Apart from the Grid activities, he is also the Executive Officer of Pacific Neighbourhood Consortium (PNC), Project Leader of International Collaboration for the National Digital Archive Program II (NDAP II) of Taiwan and the Founding President of Software Liberty Association of Taiwan (SLAT) among many other organizations and committees. He has overseen projects in several major areas at Academia Sinica. In 1996, he built the first large scalable PC Farm in Taiwan in 1996 with hundreds of processing units; while in 1997, he built the Taipei GigaPoP dark-fibre infrastructure and Taiwan's second-generation Research/Education international backbone from T1 to T3. He also pioneered the Digital Library/Museum Pilot Project in Academia Sinica which later led to the National Digital Archive Program. He received his Ph.D. degree from Edinburgh University in Theoretical Physics. His current research interests include Grid Computing, Computational Physics, Statistical Physics and Field theory, Metadata and Digital Archives. He is also adjunct professor in several universities.

Ying-Ta Wu received Ph.D from State University of New York at Buffalo. He is currently assistant research specialist of Genomic Research Center, Academia Sinica, Taipei. His research is devote to facilitate technology for probing hot-area that link to specific function in biomolecular recognition and biomolecule-compound interactions, to assist research PIs in defining features of targets for drug design, and to devise strategies for structure-based drug discovery.