Ecological Network-Based Grid Middleware for Bioinformatics Applications

L. Gao[1], L.-H Ren[1], Y.-S. Ding[1, 2, *], H. Dai[1], and S.-H. Shao[1, 2]

[1] College of Information Sciences and Technology
[2] Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education
Donghua University, Shanghai 200051, P. R. China
* Email: ysding@dhu.edu.cn

Abstract

We report on our experiences of using Ecological Network-based Grid Middleware to build a prototype system for bioinformatics applications in large-scale grid environments. The system architecture and some key technologies of the prototype are demonstrated. The implementation of remote data retrieval in the biological databases using an agent asynchronous migration approach, quantitative model management using modelbase approach, and collaborative content sharing among multiple terminal devices are focused on in the proposed prototype system.

Keywords: grid middleware, bioinformatics, ecological networks, agent asynchronous migration, modelbase system, collaborative content sharing

1. Introduction

Currently, the accumulation speed of biology related data has greatly exceeded people's expectations. What matters more principally is the complexity involved in using the data, which makes the scientists feel difficult in the bioinformatics research. To address this challenge, integrating the network infrastructure, math models, data depositories, and software tools of collecting, storing, and processing various biological information may be an important means.

Grid technologies provide the feasibility of this kind of integration. They have emerged to effectively link the heterogeneous systems, enable large-scale flexible resource sharing and innovative applications [1]. Grid technologies, initially focused on orchestrating computational resources for expensive and complex analysis, are evolving into an on-demand seamless and dynamic integrator of any kind of resources (e.g., sensors, databases, and applications). Grid technologies have been utilized for the bioinformatics applications, where myGRID project [2] is a typical instance to provide open source grid middleware to enable a virtual workbench for data-intensive bioinformatics. Another famous application is DataGrid [3], which is a pioneer in identifying the biomedical imaging field as an application domain that can benefit from grid technologies.

Our work seeks to explore the use of Ecological Network-based Grid Middleware (ENGM) [4] to set up a large-scale bioinformatics service platform, in order to integrate the effective data, models and online devices for providing high efficient, easy-to-use, and special services to the bioinformatics research. In our previous work, we developed the ENGM, a *natural ecosystems*-inspired multi-agent grid system, where agents may be very simple but their collective behaviors and the overall functionality arising from their interactions exceeds the capacities of any individual agent. The middleware achieves built-in mechanisms of self-organization, survivability, and self-evolution. We expect it can deal with the complexity involved in integrations.

In this paper, we report on our experiences of using ENGM to build a prototype system for bioinformatics applications. The prototype system currently focuses on implementation of remote data retrieval in the biology databases (e.g. Genbank [5]) using an agent asynchronous migration approach, quantitative model management using modelbase approach, and collaborative content sharing among multiple terminal devices. Due to the flexibility and the bright future of the mobile computing environment, supporting to it in the prototype system is also in our range of discussion.

The rest of the paper is organized as follows. Section 2 presents the architecture of the prototype system for bioinformatics applications based on ENGM. The key technologies and the current implementation status of the prototype system are demonstrated in Section 3. Finally, Section 4 concludes our research efforts.

2. The Architecture

We have developed a prototype system of ENGM for bioinformatics applications. Its architecture is shown

in Fig. 1. Four layers are constructed and each layer has its isolated functions.

The first layer includes all kinds of basic resources, such as bioinformatics database, models, service support servers and so on. Service support servers can provide storage capabilities and computing power for mobile devices. Agents can access those resources.

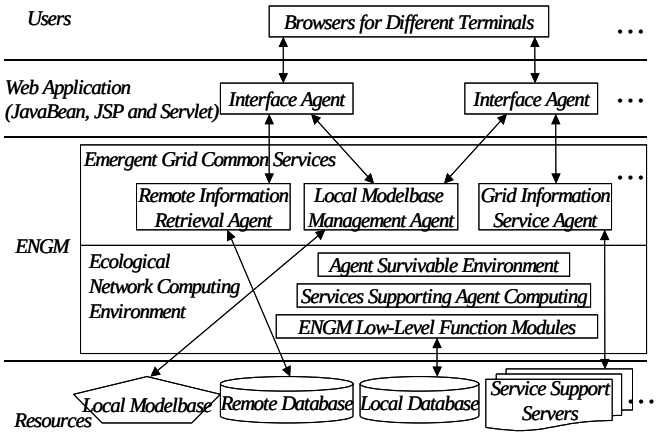

Fig. 1: The architecture of the prototype system.

The second layer is the ENGM platform, which is the core of the prototype system. It can be divided into two sub-layers: ecological network computing environment and emergent grid common services. (a) Ecological network computing environment includes ENGM low-level function modules, services supporting agent computing, and agent survivable environment. It provides a new computing and problem-solving paradigm by combining natural ecosystem mechanisms with agent technologies, supporting agent mental state representation, negotiation, communication, and cooperation based on ecological laws. Agent survivable environment is the runtime environment for deploying and executing agents. Services supporting agent computing provides a set of general-purpose runtime services that are frequently used by agents, such as naming service, energy-driven control service, evolution state management service, and agent migration service. These services alleviate agents from low-level operations and allow agents to be lightweight by separating them from the routine work. ENGM low-level functional modules are responsible for low-level operations such as the transport of messages. (b) Emergent grid common services are responsible for resource location and allocation, authentication, information service, task assignment, and so on. Note that these common services emerge from those agents and their invoking low-layered services. Remote Information Retrieval Agent (RIRA), Local Modelbase Management Agent (LMMA) and Grid Information Service Agent (GISA) are included in this sub-layer.

The third one is web application layer that bases on the web and establishes a friendly interface between users and the prototype system. Many services in this layer are provided through web pages and can be implemented by JavaBean, JSP, Servlet, and so on.

In the fourth layer, an interface agent will be generated corresponding to the user when a user login this web. Also the interface agent will register in the ENGM platform. If the user wants to search for remote bioinformatics data, the query will be generated by the interface agent and be sent to a RIRA. After process the query, the RIRA will send the search results to the interface agent. Finally, the user can get the results from web page that is generated by the interface agent.

Note that the livings and actions of all agents in ENGM depend on energy, which is an important concept inspired from ecosystem. Agents must store and consume energy for their livings. They expend energy for their usage of resources. To gain energy from users or other agents, an agent may perform a service. Viewed from this, energy is similar to money in economic world. In addition, the changes of agent actions and the transitions of agent evolution states are driven by energy. For instance, the more abundant energy an agent has, the higher demand it needs. If the energy expenditure of an agent is more than its energy earning by providing a service, the agent will lack of energy and lose the permission to use resources. As thus, it dies from its wasteful energy performances.

3. System Service Implementation

3.1. Agent Asynchronous Migration

The prototype system use migration mechanism of mobile agents [6] to implement remote data retrieval of Genbank. The principle is according to record format of Genbank is relatively unvarying, utilizing codes to seize effective information from the web tags. The ability of migration provides mobile agents a means to overcome the high latency or limited bandwidth problem by moving their computations to required resources or services. Considering that the system services should support mobile computing environments, the proposed system has offered two migration modes (synchronous and asynchronous) in the runtime services of agent survivable environment. To mobile computing environments, agent asynchronous migration plays an even more important role: not only overcomes the resource limitation of mobile terminals, but also does not need mobile users keep persistent connection to Internet. Using

asynchronous migration approach, RIRAs can work well even if users are in the low-quality connection or disconnection status in their most time.

To agent asynchronous migration, the migration request is not dealt with immediately, but inserted into the processing queue at first. The flow chart is shown in Fig. 2, when the request is under processing, the migration service of ENGM checks the energy information in the source node to decide to migrate or not. In addition, the migration service still judge if the target node accepts the migration, the target node is reachable, and the migration is successful.

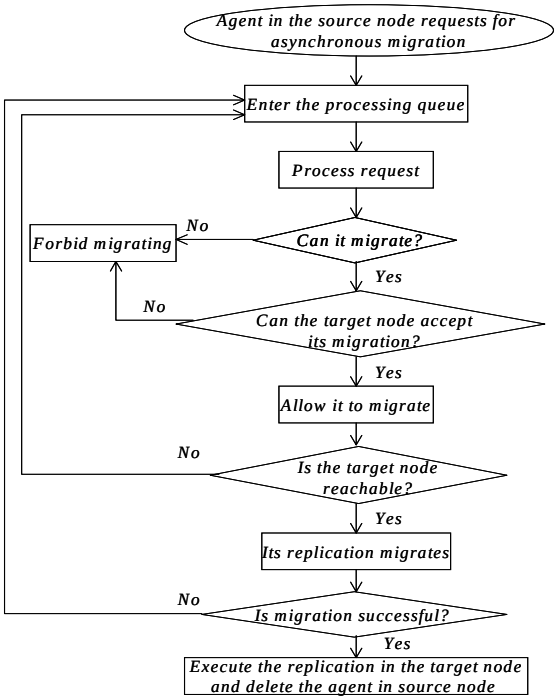

Fig. 2: The flow chart of agent asynchronous migration.

The RIRA need defray much energy for the above services. If the user's request is delayed due to the excessive checking, it will be punished by decreasing its energy. Thus, the control mechanism forms to make RIRAs be responsible for their behaviors.

3.2. Model Management

Gene sequence analysis is a complex problem. Generally, making an important conclusion should depend on the synthesis and integration of many analysis models. There are many models in common use such as direct repeats model, Needleman-Wunsch algorithm model, and so on. In addition, to carry on more accurate analysis and make comprehensive judgment, more considerations such as protein structure and related gene sequences are needed.

In the prototype system, we use modelbase approach [7] to implement the combination execution of multiple quantitative models. LMMAs in the ENGM platform are responsible for carrying the management into execution. Fig. 3 gives out an example of using this technology to implement our developed SARS-CoV sequence analysis model [8]. The model is a combination model including a few sub models. When it is invoked, the invoking list of sub models is generated. *Sequence alignment*, *statistics of base distribution*, *Z curve* and *coordination model* in Fig. 3 are combination options. One of the corresponding result displays is shown in Fig. 4.

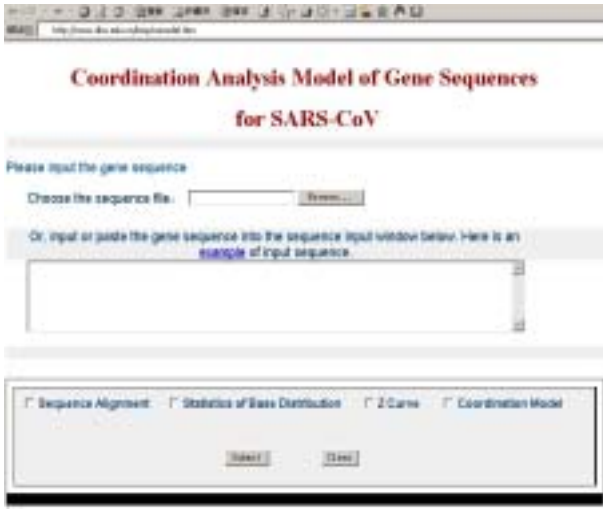

Fig. 3: The implementation of SARS-CoV sequence analysis model based on combination models.

3.3. Collaborative Content Sharing among Multiple Devices

To enable a large-scale and pervasive resource sharing, the prototype system should consider the union of mobile multi-device computing and collaborative multi-user computing. Three components are needed in collaborative content sharing: service support servers, GISAs, and interface agents.

Service support servers store the multi-media contents and a sharing strategy file that defines different contents sent to different users and devices. The contents and the strategy file are described using XML so that agents can understand them easily.

GISAs can seize and parse the contents and the strategy file. According to the unified XML sharing strategy file, it splits the requested contents and delivers the appropriate partial view of contents to each interface agent.

Interface agents can contact each user (user agent) who requests service contents for the information on the user's preferences and device characteristics, and

generate personalized partial views among the devices available to each user.

We evaluate the service in a simulation environment where different mobile phone emulators running on different PC acts as different device in the real world. The content we want to share collaboratively in multiple devices is the result of SARS-CoV sequence analysis as shown in Fig. 4. Assuming that different users have different requirements, we design a strategy file. Interface agents use XSLT (Extensible Stylesheet Language Transformation) to transform the content format into suitable formats such as WML (Wireless Markup Language). Fig. 5 demonstrates the simulation results of J2ME mobile phone emulator, WAP PDA emulator, WAP mobile phone emulator and Motorola i85s emulator from (a) to (d).

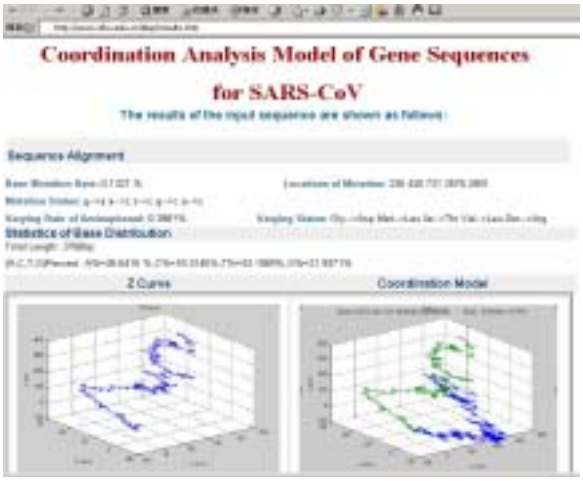

Fig. 4: The results of SARS-CoV sequence analysis.

(a) (b) (c) (d)

Fig. 5: The simulation of collaborative content sharing among different emulators.

4. Conclusion

We set up a bioinformatics service-oriented prototype system using ENGM. Considering the instable network connection, agent asynchronous migration flow is proposed to conduct the remote retrieval of the biology related data. The modelbase approach is used to make the invoking and combination of bioinformatics calculation models convenient. In addition, an approach dealing with collaborative content sharing among multiple users and devices is given out. An application example of using these two approaches to implement SARS-CoV sequence analysis is also demonstrated.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (No. 60474037 and 60004006), Program for New Century Excellent Talents in University, and Specialized Research Fund for the Doctoral Program of Higher Education from Educational Committee of China (No. 20030255009).

References

[1] I. Foster, Internet Computing and the Emerging Grid, *Nature*, 7 Dec. 2000, http://www.nature.com/nature/webmatters/grid/grid.html.

[2] C. Wroe, C. A. Goble, M. Greenwood, P. Lord, S. Miles, J. Papay, T. Payne, and L. Moreau, "Automating Experiments Using Semantic Data on A Bioinformatics Grid", *IEEE Intelligent Systems*, 2004, 19(1): 48-55.

[3] J. Montagnat, F. Bellet, H. Benoit-Cattin, V. Breton, L. Brunie, H. Duque, Y. Legré, I.E. Magnin, L. Maigne, S. Miguet, J.-M. Pierson, . Seitz, and T. Tweed, "Medical Images Simulation, Storage, and Processing on the European DataGrid Testbed", *J. Grid Computing*, in press.

[4] L. Gao, Y.-S. Ding, and L.-H. Ren, "A Novel Ecological Network-Based Computation Platform as Grid Middleware System", *Int. J. Intelligent Systems*, 2004, 19(10): 859-884.

[5] National Center for Biotechnology Information Genbank, http://www.ncbi.nlm.nih.gov/Genbank.

[6] W. S. E. Chen and C.-L. Hu, "A Mobile Agent-Based Active Network Architecture for Intelligent Network Control", *Information Sciences*, 2002, 141(1-2): 3–35.

[7] M. Hirafuji, K. Tanaka, T. Kiura, A. Otuka, "Modelbase System: A Distributed Model Database on the Internet", *Int. Workshop on Asia Pacific Advanced Network and its Applications*, pp. 57-61, 2000.

[8] L. Gao, Y.-S. Ding, H. Dai, Z.-D. Huang, S.-H. Shao, "A Novel Fingerprint Map of SARS-CoV with Visualization Analysis", *the Third International Conference on Image and Graphics*, pp. 226-229 , 2004.