A SCALABLE FRAMEWORK AND PROTOTYPE FOR CAS E-SCIENCE

Yuanchun Zhou*, Yun Xiao, Kaichao Wu, and Baoping Yan

*****Computer Network Information Center, Chinese Academy of Sciences, Beijing, China*

Email: yczhou@sdb.cnic.cn

ABSTRACT

Based on the Small-World model of CAS e-Science and the power low of Internet, this paper presents a scalable CAS e-Science Grid framework based on virtual region called Virtual Region Grid Framework (VRGF). VRGF takes virtual region and layer as logic manage-unit. In VRGF, the mode of intra-virtual region is pure P2P, and the model of inter-virtual region is centralized. Therefore, VRGF is decentralized framework with some P2P properties. Further more, VRGF is able to achieve satisfactory performance on resource organizing and locating at a small cost, and is well adapted to the complicated and dynamic features of scientific collaborations. We have implemented a demonstration VRGF based Grid prototype—SDG.

Keywords: e-Science, Grid, P2P, Small-World, Framework

1 INTRODUCTION

"E-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it" (Taylor, 2002). E-Science enables scientists to generate, analyze, share, and discuss their insights, experiments, and results in a more effective manner. These experiments involve geographically distributed and heterogeneous resources such as computational resources, scientific instruments, databases, and applications. The data in these experiments are usually massive and distributed across numerous institutions for various reasons including: the inherent distribution of data sources; large-scale storage and computational requirements; the need to ensure high-availability and fault tolerance of data; and caching to provide faster access. CAS (the Chinese Academy of Sciences) e-Science (Nan, 2002) is built upon the mass scientific data resources of the Scientific Database (SDB), in which multi-disciplinary scientific data are accumulated through the course of scientific activities in the CAS. The nodes in CAS e-Science are located at various institutes, of which each has a specific research domain. Scientific activities in the CAS have Small-World characteristics. For example, in the scientific collaborations graph, the nodes are in research institutes, and two research institutes are connected if they have the same service properties and research domains. The vision of CAS e-Science is to take valuable data resources into full play by benefiting from advanced information technologies, Grid technology and P2P technology in particular.

The representatives of Grid systems, such as Globus (Foster, 1997) and Web services (W3C, 2003), provide flexible and uniform access interfaces to all types of resources with grid service or web service and are used to build the grid application. However, in these systems, the computing mode is client/server, and the services are published and discovered with a centralized mode, which has poor scalability and a single concentrated point of failure. Some P2P systems, e.g. Napster (Napster website), are distributed file systems with centralized modes, which manage the directory by a server. They provide strong capability to manage open file-sharing systems**.** However, they still use a centralized mode, so they have a single point of failure. Other P2P systems, such as Gnutella (Gnutella website) (Zeinalipour, 2002) and Freenet (Clarke, 2002), are pure P2P systems, which have a completely decentralized structure. They have the great advantage of having no bottlenecks and good robustness. However, they are faced with some challenges such as security, network bandwidth, and architectural design and are difficult to search for services that are clustered together as described by WSDL (Christensen, 2001) and GSDL (Foster, 2002). They also have poor scalability and low efficiency because of the exponential growth of redundant information. VDHA (Huang, 2002) is a virtual and dynamic hierarchical architecture, in which Grid nodes are grouped virtually. It has a decentralized architecture with some P2P properties and also has scalable, autonomous, exact, and full service

discovery properties. However, it lacks a mathematic model, and the classifying strategy of the virtual organization is not clear. The Sun's Project JXTA 2.0 Super-Peer Virtual Network (Traversat, 2004) is similar to VDHA.

We use the model of "Small-World and power law" as a theoretical foundation and present the grid framework based on virtual regions called Virtual Region Grid Framework (VRGF). This framework is decentralized and scalable and implements a scalable VRGF based Grid system—SDG (Scientific Data Grid), which combines the advantages of P2P and C/S.

The rest of the paper is organized as follows. In Section 2, the VRGF model and its related protocols are described. Section 3 lays out an implementation prototype of VRGF. Finally, we give the conclusion and outline future work.

2 AN OVERVIEW OF VRGF

2.1 Definition and Related Concepts

Definition 1 (Power Law) - *the distribution probability of the nodes with* k *degree in a random graph is expressed as* $P(k) \propto k^{-\tau}$, *where* $1 < \tau < \infty$. *In a network, a few nodes are of a high degree, and many nodes are of a low degree. Therefore, there is a high probability of finding related information through the high degree nodes* (Faloutsos, 1999).

Definition 2 (the Small-World) - *the Small-World is a network topology that has a large clustering coefficient and small average path length* (Kleinberg, 2000).

Definition 3 (Service Similarity) - *the service similarity of two nodes is the similarity of the service properties and the research domains between the two nodes. The function of service similarity between two nodes is expressed as* $Sim(N, N') = \xi_1 Sim_d(N, N') + \xi_2 Sim_p(N, N')$, $\sum_i \xi_i = 1$, $0 \le \xi_i \le 1$, $i = 1,2$, *where* ξ_1 *and* ξ_2 *are weights,* $Sim_p(N, N')$ *is the similar function of the service properties.* $Sim_d(N, N')$, *which is the similar function of the research domains, can be defined as* $Sim_d(N, N') = \mu_1 Sim_{name}(N, N') + \mu_2 Sim_{text}(N, N')$, $\sum_i \mu_i = 1$, $0 \le \mu_i \le 1$, $i = 1,2$, *where* μ_1 *and* μ_2 *are weights.* $Sim_{name}(N, N')$ *is the similar function based on domain name.* $Sim_p(N, N')$ *is the similar function based on domain text. The match algorithm based on the keywords is adopted in* $Sim_p(N, N')$, $Sim_{name}(N, N')$ *and* $Sim_{text}(N, N')$. *For example,* $Sim_{name}(ICT, ISCAS)$ *is 1 because ICT and ISCAS have the same research domain name.*

Definition 4 (Host) - *a client host is an apparatus (such as desktop computer, PDA, mobile computer, etc), which is used to log into a Grid system.*

Definition 5 (Virtual Region) - *a Virtual Region is formed virtually by Grid nodes based on the service properties and the research domains of the nodes. The nodes of intra-virtual regions have high similarity.*

Definition 6 (Grid Node) - *a grid node is an ordinary node in the Grid system.*

Definition 7 (Head Node) - *a head node is a Grid node which manages the virtual region. The highest performance node in the virtual region is chosen as the head node, which locates the logical center of the virtual region. The head node provides the properties of the virtual region and the interface for the inter-virtual region.*

Definition 8 (Active Grid Node) - *an active grid node is the Grid node that takes charge of a node joining to the Grid. Each virtual region has an active grid node. In this framework, we use the head node as the active grid node. A node can take any active grid node as an entrance node to join to the Grid system.*

Definition 9 (Layer) - *the virtual regions of Layer* L_i *are composed of the head nodes of* L_{i-1}.

2.2 The Description of VRGF

The nodes of CAS e-Science are usually located in institutes. The institutes are always formed into virtual regions according to specific domains, and several virtual regions share a more general common domain. This is similar to Small-World networks (Duncan, 1999). Two characteristics distinguish Small-World networks: first, a small average path length, typical of random graphs; second, a large clustering coefficient that is independent of network size. The clustering coefficient means the number of a node's neighbors that are connected to each other. One can picture a Small-World as a graph constructed by loosely connecting a set of almost complete subgraphs. Thus, CAS e-Science has the Small-World property. The Small-World example of scientific activity is the scientific collaboration graph, where the nodes are scientists, and two scientists are connected if they have the same research field. Such graphs with a Small-World character in scientific collaborations can span a variety of different domains, including physics, biomedical research, mathematics, and computer science. Then, CAS e-Science, in which nodes are institutes and edges are relationships among institutes having the same research domains, becomes the Grid system with the Small-World character.

According to the Small-World model of CAS e-Science and the power law of the Internet, the nodes in VRGF are formed virtually to virtual regions based on the service properties and the research domains. Virtual regions are virtually hierarchical, with one root-layer, several middle-layers, and the lowest layer (layer 0). Figure 1 shows the network topology of VRGF. The network topology has many layers, and each layer is composed of virtual regions, which include many Grid nodes. Any node of the Grid system belongs to one or more virtual regions. All physical nodes are in layer 0 virtual regions. Among these nodes of layer 0, one (just one) node (called the head node) in each virtual region is chosen to form the upper-layer (layer 1) virtual region. From the nodes in these upper-layer virtual regions, one is chosen to form the upper-upper-layer (layer 2) virtual region in the same way, and this is repeated until one root-layer (only one node) is formed. In the virtual region the node with the highest performance is chosen to be the head node (also the active grid node), which is not only in the low-layer, but also in the upper-layer. All active grid nodes in each layer are connected. Nodes can join and leave a virtual region dynamically.

In VRGF, the intra-virtual region nodes have high service similarity, and the inter-virtual region nodes have low service similarity. So there is a higher probability of satisfying specific service requests in an intra-virtual region than in an inter-virtual region and for locating services from all nodes to the intra-virtual region nodes.

Thus, VRGF topology has several properties: (1) high performance for locating services and avoiding request flooding efficiently; (2) high scalability and robustness; (3) transforming data streams to the active virtual region easily.

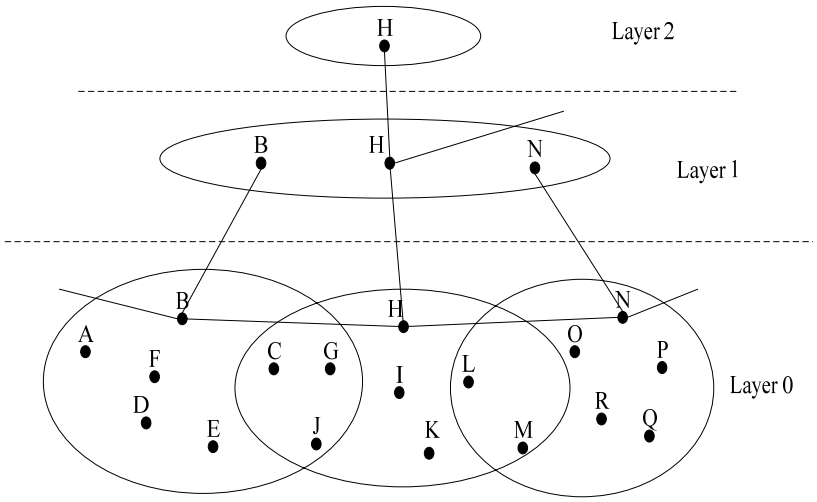

Figure 1. VRGF network topology

2.3 Virtual Region Management Protocol (VRMP)

VRMP is a protocol used to manage the virtual region and the membership of the virtual region. Because the mode of an intra-virtual region is pure P2P and the performance of locating grid services depends on the size of the virtual region, it is necessary to split and merge virtual regions. If the size of a virtual region is too large, VRGF can not avoid request flooding. If the size of a virtual region is too small, VRGF may become a complete C/S model. Therefore, splitting and merging virtual regions reasonably locates services, accesses data, and shares resources at a lower cost. Moreover, operating a virtual region occurs only in virtual regions of layer 0 because the physical nodes are all in layer 0.

The details of the VRMP algorithm are shown in the following:

while(true) {
  switch(event){
    case: virtual region R will be split
      H = Head (R); */*Get the head node of R */*
      $\{R_1, R_2\}$ =Divide (R); */*Divide R in two, delete all edges connecting R_1 and R_2, assuming R_2 including H */*
      H_1 =Head_Choise (R_1); */*Chose new head node in virtual region R_1 */*
      Update (H); */*Update H information, such as the number of nodes*/*
      Update (H_1); */*Update H_1 information, such as the number of nodes*/*
      Put_UpLayer (H_1 , H); */*Put H_1 into virtual region of up-layer, which includes H */*
      VR = Get_UpVR(H); */*Get upper-layer virtual region which includes H */*
      Update (Head (VR))
  case: virtual region R_1 and R_2 will be merged
      H_1 = Head (R_1); H_2 = Head (R_2); */*Get the head nodes of R_1 and R_2 */*
      H = Choise (H_1 , H_2); */*Choise the new head node from H_1 and H_2 */*
      R =Merge (R_1 , R_2); */*Merge R_1 and R_2 into R */*
      Set_Head (R , H); */*Set H to the head node of R */*
      Update (H);
      Put_UpLayer (H , H_1)
      Delete_UpNode (H_1); */*Delete node H_1 in the upper-layer*/*
      Delete_UpNode (H_2);*/*Delete node H_2 in the upper-layer*/*
      VR = Get_UpVR(H); */*Get upper-layer virtual region which includes H*/*
      Update (Head (VR))
  case: a node N joins a VRGF
      VRL_j =Find_Layer (*AGN*, -); */* Finding Virtual Regions of Root-Layer(AGN is any active grid node)*/*
      For $i = 1$ to $|P|$ */*$|P|$ is the number of elements*/*
        While ($j > 0$) */* parallel execution*/*
          Find M s.t Maximize $Sim(N, M)$, where $M \in VRL_j$
          VRL_{j-1} =Find_Layer (M , $j-1$); */*Finding virtual region of in Layer $j-1$ which M belongs to*/*
          $j--$;

$VRL_j = VRL_{j-1}(M)$;
Endwhile
JoinVirtualRegion VRL_j;
Endfor
case: a node N leaves a VRGF
$|VRSet|$ = Get_VR (N , 0); */*Get the layer 0 virtual region set which N belongs to*/*
For $i = 1$ to $|VRSet|$ /* *parallel execution**/
LeaveVirtualRegin (*N*, $VRSet[i]$); */*N leave from virtual region* $VRSet[i]$ **/*
Update (Head ($VRSet[i]$);
Endfor
}
}

3 VRGF-BASED GRID PROTOTYPE

3.1 VRGF-Based Grid Prototype Architecture

The CAS e-Science project supported by CAS is aimed at enhancing science and technology research by virtual cooperation via the Internet. There are now 42 research institutes, each having a special domain. In order to combine these 42 research institutes into the CAS e-Science System, we use VRGF to model the CAS e-Science Grid System prototype (called SDG). All research institutes are turned into many virtual regions in layer 0 based on a common research domain. For example, ICT, ISCAS, and CNIC are the nodes of the network virtual region because they have a common research domain: networks. According to the more general common research domain, the virtual regions in layer 0 are formed into layer 1. For instance, the computer virtual region in layer 2 includes the network virtual region, the software virtual region, the hardware virtual region, etc. The root-layer has one node with the most generally common research domain (see Figure 2).

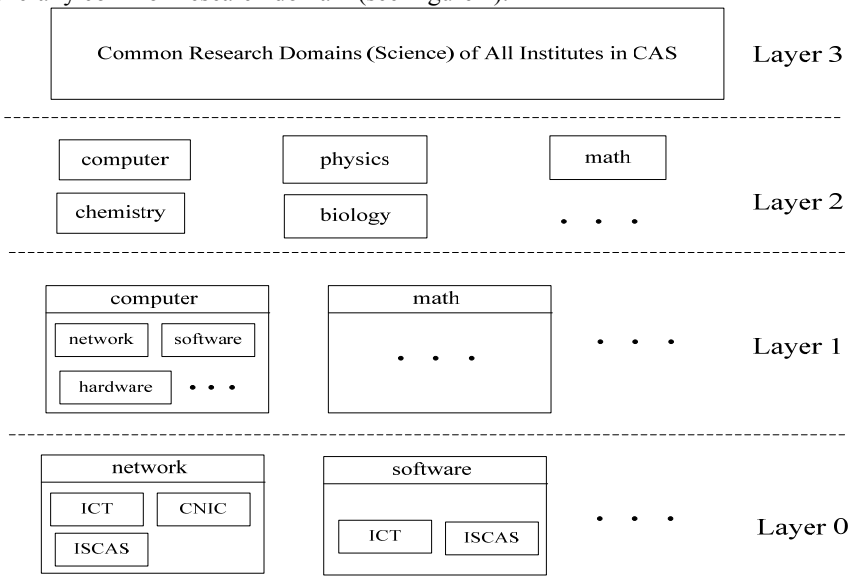

Figure 2. VRGF architecture of SDG

3.2 Message-Based Implementation of VRGF

The implementation of VRGF is based on the message/event, as Figure 3 shows. The main design entities within the implementation are:

(1). Event: The Event is an abstract object that represents the change of another object.

(2). Message: The Message is used to describe data of this framework. We use the unified message format described by Abstract Syntax Notation One (ASN.1) (Larmouth, 2003) as follows:

```
message format:=SEQUENCE{
        Version, /*message format version */
        ID, /*the identifier of the message*/.
        Group_ID, /*the identifier of group message*/
        Source_Global_Entity /* the entity who sends the message */
        Destination_Global_Entity, /* the entity who receives the message */
        Type, /* the type of message */
        Transport_Protocol, /* SOAP or XML, etc. */
        Port, /* port number */
        Length, /*the length of the message*/
        Message_Body /* the message content */
}
```

As the message format indicates the message type and protocol, the receiver can explain the message content correctly.

(3). Task: A task is the fundamental unit of work in our framework. It is a typed message that contains the description of some work to be done, along with the data required to complete that task. Tasks are processed in a series of stages by individual components of the application. Stages of a task can either be executed in sequence, in parallel, or in a combination of the two. By decomposing the task into a series of stages, it is possible to distribute those stages over multiple physical resources and allocate those tasks for load-balancing and fault-isolation purpose.

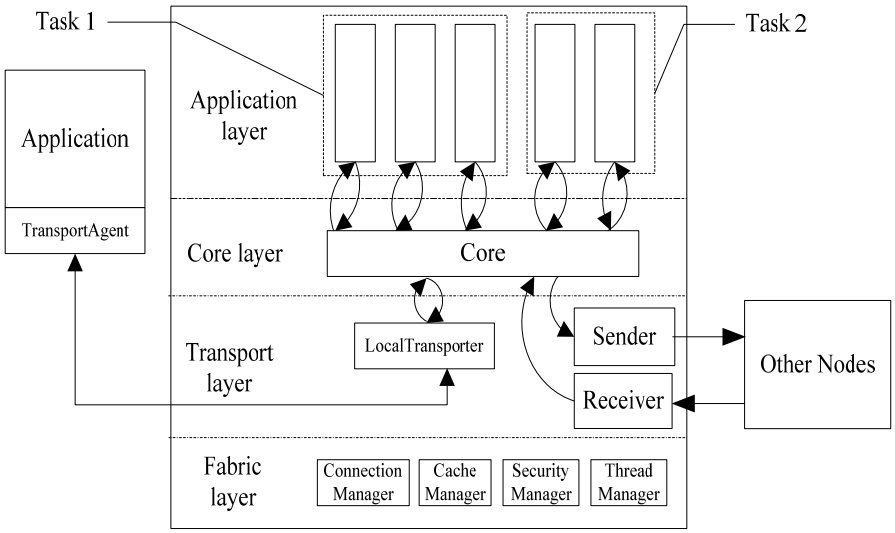

Figure 3. The message-based implementation of VRGF

The implementation architecture includes four layers. The bottom layer is a fabric layer, which provides the fundamental functions, such as security management, etc. The next layer is a transport layer, which takes the responsibility for communication between nodes. It contains many transport protocols, such as SOAP, XML, etc. The core layer is the next layer, which deals with the messages received from or sent to the transport layer. The top layer is the application layer which deals with the client and server-side tasks or applications.

One of the working scenarios is as follows: The client requests to query data in the Grid by sending a query message (task), which indicates the service name, searching model, domain knowledge, and so on, to the head node. Then, according to the service similarity, the head node locates the related virtual regions, in which the core layer of their head nodes decomposes the query message and dispatches to the nodes of the relevant virtual region. After querying the data of these nodes, they send the response message to the head node, which sends the result to the client.

4 CONCLUSIONS AND FUTURE WORK

VRGF adopts the Small-World model and power law to organize virtual regions based on service properties and research domains, so the nodes in an intra-virtual region have high service similarity. Therefore, VRGF has built on a mathematic model.

VRGF can solve scaling and autonomy problems and has high performance and accurate discovery of resources and services. It has a high probability of satisfying a specific service request in a virtual region, and the complexity of locating services is reduced from all nodes of the Grid to the nodes of an intra-virtual region. We have implemented a demonstration prototype.

Our further work will focus on completing the SDG, on enriching the services of SDG, and on implementing the mechanism for locating services.

5 ACKNOWLEDGEMENTS

We would like to thank our colleagues and graduate students in our Laboratory for their discussion, cooperation, and contribution. The work reported in this paper is supported by the National High Technology Research and Development Program of China (863) and by the information constructing projects of the Chinese Academy of Sciences (CAS). We are grateful for these supports.

6 REFERENCES

Christensen,E., Curbera, F., Meredith, G. ,& Weerawarana, S. (2001) Web Services Description Language (WSDL) 1.1 *W3C.Note* 15. Retrieved March 15, 2006 from the WWW: http:/www.w3.org/TR/wsdl:

Clarke, I., Sandberg, O., Wiley, B., & Hong, T. (2000) Freenet: A distributed anonymous information storage and retrieval system. *In: ICSI Workshop on Design Issues in Anonymity and Unobservability* (pp.311- 320), California: Springer.

Duncan, J. W. (1999) Small-Worlds:The Dynamics of Networks between Order and Randomness. *Princeton University Press*.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999) On power-law relationships of the Internet topology. *In: Chapin L, Sterbenz JPG, Parulkar G, Turner JS, eds. Proc. of the ACM SIGCOMM'99*(pp.251-262), New York:ACM Press.

Foster, I. & Kesselman, C. (1997) Globus: A metacomputing infrastructure toolkit. *International Journal of Supercomputer Applications* 11(2), 115-128.

Foster, I. ,Kesselman, C., Nick, M.J., & Tuecke, S. (2002) The physiology of the grid: An open grid services architecture for distributed systems integration. Open Grid Service Infrastructure WG, *Global Grid Forum*, June, 2002.

Gnutella website. Retrieved March 20, 2006 from the WWW: http://www.gnutella.com.

Huang, L., Wu, Z., & Pan, Y. (2002) Virtual and dynamic hierarchical architecture for Chinese university e-Science grid. *In: Proc. of the 2002 Int'l Workshop on Grid and Cooperative Computing (GCC2002)*(pp.297−311), Publishing House of Electronics Industry.

Kleinberg, J. (2000) The small-world phenomenon: An algorithmic perspective. *Proc32nd ACM.Symposium on Theory of Computing* (pp.820-828).

Larmouth, J. (2003) ASN.1 Complete, Retrieved April 06, 2006 from http://www.oss.com/asn1/larmouth.html.

Nan, K. & Yan, B.P. (2002) Introduction to Scientific Data Grid. *APAN Grid Workshop*, Shanghai, China.

Napster website. Retrieved March 21, 2006 from http://www.napster.com.

Taylor, J. (2002) e-Science definitions. Retrieved March 28, 2006 from http://www.e-science.clrc.ac.uk.

Traversat, B., Arora, A., Abdelaziz, M. ,Duigou, M., Haywood, C., Hugly, J.C., Pouyoul, E., & Yeager, B. (2004) Project JXTA 2.0 Super-Peer Virtual Network. http://www.jxta.org/project/www/docs/JXTA2.0protocols1.pdf/.

W3C. (2003) Web Services Architecture. Retrieved April 10, 2006 from http://www.w3.org/TR/2003/WD-ws-arch-20030808/.

Zeinalipour, Y.D. & Folias,T.A. (2002) quantitative analysis of the gnutella network traffic. *Technical report*, University of California, Riverside.