支持 e-Science 的网格体系结构及原型研究*

黄理灿+, 吴朝晖, 潘云鹤

(浙江大学 计算机科学与技术学院,浙江 杭州 310027)

A Grid Architecture for Scalable e-Science and Its Prototype

HUANG Li-Can+, WU Zhao-Hui, PAN Yun-He

(College of Computer Science, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: Phn: +86-571-87951647, E-mail: lchuang@cs.zju.edu.cn, lchyuang@hzcnc.com, http://www.zju.edu.cn

Received 2003-02-19; Accepted 2003-10-08

Huang LC, Wu ZH, Pan YH. A grid architecture for scalable e-Science and its prototype. *Journal of Software*, 2005,16(4):577−586. DOI: 10.1360/jos160577

Abstract: This paper presents an e-Science Grid architecture called Virtual and Dynamic Hierarchical Architecture (VDHA). VDHA is a decentralized architecture with some P2P properties. It also has scalable, autonomous, exact, and full service discovery properties. We have implemented a demonstration VDHA_based Grid prototype — VDHA_Grid, which has scalable Grid information service. VDHA_Grid is the core software for the project of the Chinese University e-Science Grid. In this paper, advantages and several protocols of the VDHA are also discussed.

Key words: VDHA; grid; e-Science; protocol; VDHA_Grid

摘 要: 提出了 e-Science 网格的虚拟动态分层体系结构(VDHA).VDHA 是具有 P2P 特征的分散的支持 e-Science 网格的体系结构.VDHA 具有可扩展性、自动性、精确和完全服务发现的特点.实现了基于 VDHA 的验证原型系统 VDHA_Grid,其具有可扩展的网格信息服务.VDHA_Grid 是中国大学 e-Science 网格项目的核心软件.也讨论了 VDHA 的一些优点及相关协议.

关键词: VDHA;网格;e-Science;协议;VDHA_Grid

中图法分类号: TP393 **文献标识码**: A

1 Introduction

"e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it"[1]. E-Science enables scientists to generate, analyze, share, and discuss their insights, experiments,

* Supported by the Virtual Cooperative Research Project Granted by the Ministry of Education of China (高校网上合作研究中心平台建设项目)

HUANG Li-Can got his Ph.D. degree in 2004 from Zhejiang University. Now he is a Senior Engineer at College of Computer Science and Technology, Zhejiang University. **WU Zhao-Hui** is a professor at College of Computer Science and Technology, Zhejiang University. **PAN Yun-He** is a professor at Zhejiang University and an academician of the Chinese Academy of Engineering.

© 中国科学院软件研究所 http://www.jos.org.cn

and results in a more effective manner. The main characteristics of e-Science are the coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations called VOs[2], and dynamically involving a large number of nodes generally distributed globally in geography.

The computing architecture of e-Science is usually based on Grid[2−4]. The representatives of Grid systems are Globus[5] and Web services[6]. However, in these systems, the computing mode is client/server, and the services are published and discovered with centralized mode, which is with bad scalability and a single point of failure. P2P[7] has good scalability, but it has some challenges such as security, network bandwidth, and architecture designs, and is difficult to search services which are clustered together and described by WSDL[8] or GSDL[3]. We present a Virtual and Dynamic Hierarchical Architecture (VDHA) which is decentralized and scalable, and implement a scalable Grid system VDHA_Grid, which combines the advantages of P2P and C/S.

The structure of this paper is as follows. Section 2 describes the VDHA and related protocols. Section 3 gives an example of the virtual cooperative research projects granted by Ministry of Education of China, and finally we give conclusions.

2 Overview of VDHA

2.1 Description of VDHA

We define those kinds of Grid, whose nodes are located in the Universities or Institutes for scientific research, as e-Science Grid. The nodes are relatively stable compared with other type of Grids. Figure 1 shows the network architecture and a Grid node overlay topology for VDHA_Grid. In this network architecture, there are a core circle formed by Grid nodes, and a surrounding circle consisting of desktop computers, mobile computers, PDAs, sensors, other networks, etc. The core circle uses the virtual and dynamic hierarchical architecture—VDHA[9] as its architecture. The Grid node has a sole IP address used for its identification ID. VDHA is a virtual and dynamic hierarchical architecture in which Grid nodes are grouped virtually. Nodes can join the group and leave the group dynamically. The groups are virtually hierarchical, with one root-layer, several middle-layers, and many leaf virtual groups (these groups are called VOs). Among these nodes of VOs, one (just one) node (called gateway node) in each group is chosen to form upper-layer groups, and from the nodes of these upper-layer groups upper-upper-layer groups is formed in the same way. This way is repeated until one root-layer group is finally formed. In the same group, all nodes are capable to be a gateway node. A gateway node is the node which is not only in the low-layer group, but also in the up-layer group. Gateway nodes will forward the low-layer group's status information to all the nodes in the up-layer group, and distribute the upper-layer group’s status information to all the nodes in the lower-layer group.

2.2 Formal definition of VDHA

Definition 1. A Grid node (denoted as p) is the node in a Grid system. All of the *p*s form a set *PS*, that is, $PS=\{p_i \mid i\in N\}$, $N=\{1\ldots n\}$, here, n is the number of the Grid nodes and each p_i has an ID (usually Internet IP address).

Definition 2. An entrance node (denoted as ent) is a Grid node which is an entrance point for users to log into a Grid system.

Definition 3. An owner node (denoted as ow) is a Grid node which manages the Users.

Definition 4. A user (denoted as user) is the role which uses a Grid system. User is managed only by its owner node, not by the entire Grid. And it may be the same user which belongs to an owner node before the owner node joins a Grid system.

© 中国科学院软件研究所 http://www.jos.org.cn

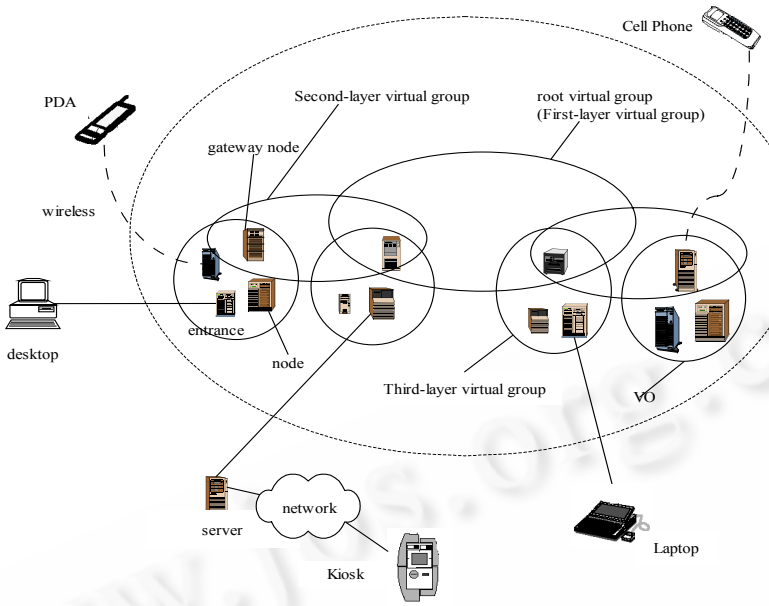


Fig.1 Network architecture of VDHA_Grid and structure of VDHA

Definition 5. A client host (denoted as cli) is an apparatus (such as desktop computer, PDA, mobile computer, etc), which is used by users to log into a Grid system and to do business.

Definition 6. A gateway node (denoted as gn) is a Grid node which takes the coordinating functions in several different layers of the virtual groups.

Definition 7. A virtual group (denoted as VG) is formed virtually by Grid nodes. VG^i_α means the group is in the i-th layer and the name of this virtual group is α. A virtual group is identified by its group name and layer number.

Definition 8. A coordinator of virtual group (denoted as cvg) is a gateway node taking the coordinating functions in a virtual group. The symbol cvg^i_α ($cvg^i_\alpha \in VG^i_\alpha$) means that it is a gateway node in the i-th layer α-named virtual group which functions as a coordinator.

Definition 9. A virtual group tree (denoted as VGT) is a hierarchical tree formed by virtual groups.

In VGT there is a root virtual group (symboled as RVG), and many leaf VGs are called virtual organization (denoted as VO). $VO^m\alpha_m$ means that the virtual organization is in the m-th layer and its name is α_m.

The order of layers is counted from RVG, which is defined as the first-layer VG.

VG except VO is formed purely by gateway nodes. VO is formed by Grid nodes with one (and just one) gateway node.

RVG can not be a VO, and VO can be within all the layers except the first layer.

$\boldsymbol{N}^i_\alpha$ is the number of the nodes in VG^i_α.

$\boldsymbol{N}^i_g$ is the number of the virtual groups in the ith-layer of VGT.

Definition 10. VDHA is a virtual group tree with depth of at least two layers. VDHA has dynamic properties in the number of Grid nodes, layers and virtual groups, virtual group compositions, and so on.

In VDHA, we have the following properties:

1. $VG^i_a=\{gn\in VG^{i+1}_\beta | \beta\in A^i\}$, $i>0$, where, A^i is the subset of the names of the i-th layer virtual Groups (This means that the VG is formed from lower-layer groups.)
2. If $gn_1\in VG^i_a \cap gn_1\in VG^{i+1}_\beta$ and $gn_2\in VG^i_a \cap gn_2\in VG^{i+1}_\beta$, then $gn_1=gn_2$.

© 中国科学院软件研究所 http://www.jos.org.cn

3. Each VG has one and only one node (cvg) which takes coordinating functions.
4. Grid node P can join more than one VO.
5. $PS=VO_1\cup VO_2\cup\ldots VO_{n1}$, where n1 is the number of virtual organization.
6. If p satisfies the following condition: $p\in VO^m\alpha_m \cap p\in VO^{m-1}\alpha_{m-1}\cap\ldots\cap p\in VO^{m-k}\alpha_{m-k}, m>=2$, then p is a gateway node. It is expressed with symbol $gn(m,k,\alpha_{m-k}\ldots\alpha_{m-1}\alpha_m)$. Here, m is the layer order of VO in VGT ($gn\in VO$), k is the number of layers in which the gateway node functions, and $\alpha_{m-k}\ldots\alpha_{m-1}\alpha_m$ are the names of the virtual groups from $VO^m\alpha_m$ to $VO^{m-k}\alpha_{m-k}$.

There are five roles for a node: general node which is a basic role of Grid node, gateway node which takes coordinating functions in several different layer virtual groups, coordinator which is a gateway node taking coordinating functions in a virtual group, entrance node which is an entrance point for users to log into the Grid system, and owner node which manages its own users and services. A node has potentially one or several roles of these five roles.

2.3 Grid group management protocol (GGMP)

GGMP is a protocol used to manage the membership of virtual group and virtual group tree. GGMP has two functions. Firstly, it manages the membership of virtual groups and the dynamic virtual group tree. Secondly, when a gateway node fails or leaves, it selects a new one with the maximum weighted value from all the on-line nodes in the group the gateway node is involved in. The Message Sequence Chart[10] of GGMP is shown in Fig.2, and the details of the algorithm are shown in the following.

```
while(true) {
  switch(event) {
   case: a VOα joins VDHA Grid system:
       /* a VO as a whole to join VDHA Grid, Root virtual group's symbol is VG^1_top */
       choose_new_cvg (gn, VOα, (gn∈VOα∩gnw=Maxium of piw, pi∈VOα and online));
       /* If piw, pjw, etc are with the same value, a random node is chosen. */
       set cvg=gn; /* cvg is the coordinator of VOα that is, Cvg∈VOα */
       cvg uses QDP protocol to find the interested parts of the structure of virtual group tree such as VG^k_β; /*Choose VG^k_β as an
          upper-layer virtual group to join*/
       cvg.send(JOIN_MESSAGE, cvg^k_β);
     if (cvg^k_β accepts the requisition) add (Pnode(cvg), VG^k_β) /* cvg can form a sub tree of virtual groups, and can join VG^k_β as a whole */
       cvg.send(state_table of VGroup(cvg)_message, UPcvg(cvg));
       UPcvg(cvg).send (state_table of VGroup(cvg)_message, p∈VGroup(UPcvg(cvg))); /*update down state table */
       UPcvg(cvg).send (state_ table of VGroup(UPcvg(cvg))_message, cvg);
       cvg.send(state_ table of VGroup(UPcvg(cvg))_message, p(VGroup(cvg));; /*update up state table */
  case: a VOα leaves from VDHA Grid system:
       gn=Pnode(TOP_gn(gn∈VOα);
       gn. Reselect_GatewayNode_Coordinator ();
       Delete VOα; // delete VOα
       gn=Pnode(BOTTOM_gn(gn∈VOα);
       gn.Up_Update ();
       gn=Pnode(TOP_gn(gn∈VOα));
       gn.Down_Update ();
  case: gn leaves VDHA Grid system:
       VG=VGroup(BOTTOM_gn(gn))
       gn. Reselect_GatewayNode_Coordinator ();
       set new gn1=cvg∈VG;
       gn=Pnode(BOTTOM_gn(gn∈VG));
       gn.Up_Update ();
       gn=Pnode(TOP_gn(gn∈VG);
       gn.Down_Update ();
   case: cvg fails to receive messages from p∈VGroup(cvg), p∈VGroup(UPcvg(cvg)) and p∈VGroup(Lowcvg(cvg)) exceeding a given
          time
       set gn=Pnode(cvg);
       VG=VGroup(BOTTOM_gn(gn))
       gn. Reselect_GatewayNode_Coordinator ();
       set new gn1=cvg∈VG;
       add (Pnode(gn),VG); /*change previous gn to an ordinary node.*/
       gn=Pnode(BOTTOM_gn(gn∈VOα));
```

© 中国科学院软件研究所 http://www.jos.org.cn

```
    gn.Up_Update ();
    gn=Pnode(TOP_gn(gn(VO_a));
    gn.Down_Update ();
 case: a node p leaves from a VO:
    p.send (REQUISITION_MESSAGE, cvg∈VO);
    if (cvg accepts the requisition) remove(p,VO);
       cvg.send(p_leaves_message, p∈VGroup(cvg)); /* update state table */
    cvg.send(p_leaves_message, UPcvg(cvg));
    UPcvg(cvg).send (p_leaves_message, p∈VGroup(UPcvg(cvg))); /*update up state table*/
     }
}
```

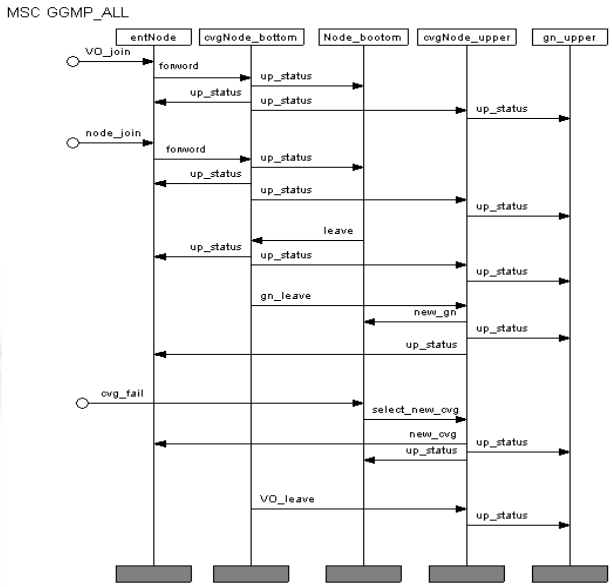


Fig.2 The message sequence chart of GGMP

2.4 VDHA_Grid architecture

Unlike Web service and Globus, VDHA_Grid publishes and registers the services into the owner node by the administrator of the node itself, and service description is based on ontology-based languages. Complexities of time, space and message-cost for service discovery of VDHA_Grid are polynomial. Figure 3 shows a VDHA_Grid architecture. In the architecture, the bottom layer is a communication layer, which takes the responsibility for the communications between nodes or between the service consumers and service providers. The upper layer is a message layer, which deals with the messages received from or sent to the communication layer. The protocols of the message layer are SOAP, ASN.1 and so on. Grid protocol layer is the next layer. It consists of many Grid protocols such as QDP, LP and so on, and core services, which is used by the Grid system itself, such as service life management service(SLMS), scalable Grid security service(SGSS), Grid monitor and control service(MCS), etc. The next layer is the service layer, which consists of the services which are consumed by the consumers and are described by the Ontology-based Grid Service Description Languages (OGSDL). The top layer is an application layer which deals with the client and server-side applications. The Security and management involves with all the above layers.

© 中国科学院软件研究所 http://www.jos.org.cn

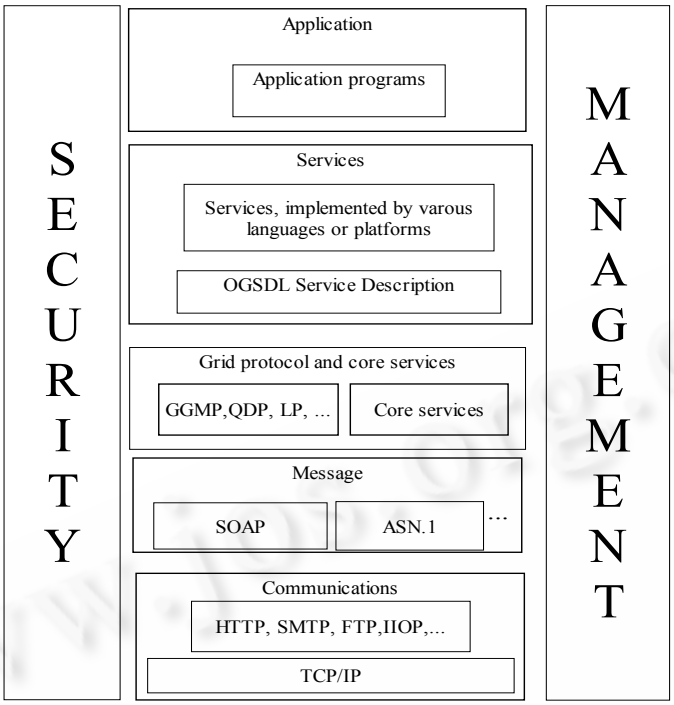


Fig.3 VDHA_Grid architecture

2.5 Query and discovery protocols

In VDHA, query and discovery protocols are used for querying and discovering some entities such as resources and services, virtual group name, node status, etc. Every node has resources and services which are described by WSDL[8] or ontology languages, etc. Matching the request message is done by the agent of node which has the services. There are two kinds of QDP: Full Search Query and Discovery Protocol (FSQDP) which searches all nodes to find nodes that match the request message, and Domain-Specific Query and Discovery Protocol (DSQDP) which searches nodes in only specific domains. FSQDP first finds out the root virtual group, and then the coordinator of the root virtual group forwards the query message to its all members. All of these members execute parallelly forwards of the message down to the members of their low-layer groups until leaf virtual groups. FSQDP has the time complexity $O(\log N)$ (N is the number of nodes), space complexity $O(N_{vg})$ (N_{vg} is the node number of each virtual group), and message-cost $O(N)$. FSQDP is effective but may cause much traffic. Domain-Specific Query and Discovery Protocol (DSQDP) (see Fig.4) is quite similar to FSQDP but it only searches the nodes whose catalogue matches the requested group keywords. To use this protocol, the object of virtual group must maintain the catalogue with the classifying services from general to detail. It may be done by the nodes joining the proper virtual group of Grid system. The protocol DSQDP has the time complexity $O(N_{vg})$, space complexity $O(N_{vg})$, and message-cost $O(N_{vg})$. This protocol is effective and the message cost is low. The details can be found in Ref.[11].

2.6 VDHA_Grid prototype implementation

The implementation of VDHA_Grid is based on the broker of message/event, as Fig.5 shows. One of the working scenarios is as following: The service consumer requests a service in the Grid by sending a query message, which indicates the service name, searching method, and so on, to the entrance node. Then QDP locates the nodes which have the service. Then Service Lifetime Management Service (SLMS) of the node creates the service

© 中国科学院软件研究所 http://www.jos.org.cn

instance, and this service instance creates other services' instances, and they provide the service functions to the consumer. After finishing providing the service, SLMS will destroy the service instance. Monitor and Control Service (MCS) is an optional service which is used to monitor and control the status of the nodes and service instances.

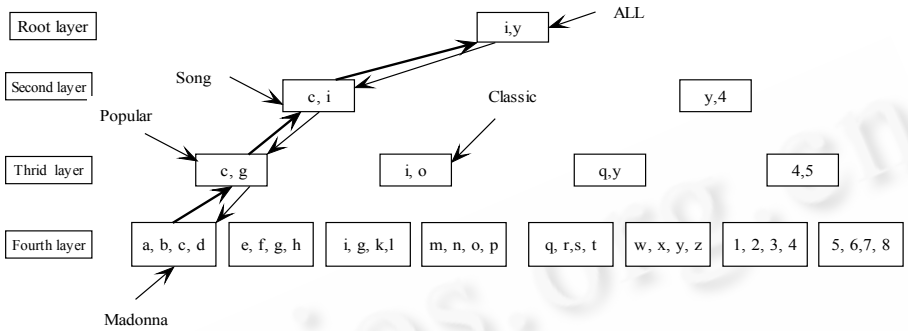


Fig.4 DSQDP searching process

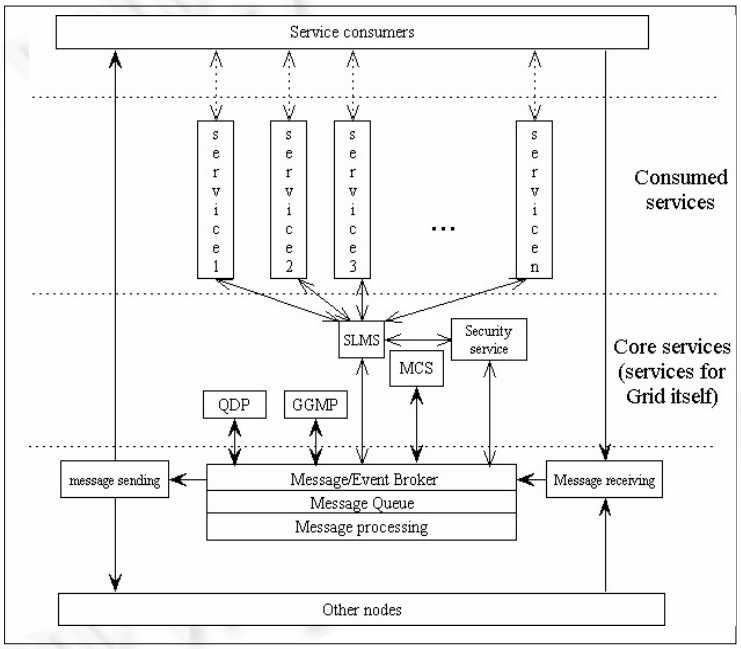


Fig.5 Message-Based implementation of VDHA_Grid

We use the unified message format described by Abstract Syntax Notation One (ASN.1)[12] as follows:

message format:=SEQUENCE{

Version, /*message format version */

Source_Global_Entity /* the entity who sends the message */

Destination_Global_Entity, /* the entity who receives the message */

Protocol, /* the protocol used to explain Message_Body*/

Transport_Protocol, /* UDP or TCP, etc. */

Port,/* port number */

Message_Body /* the message content */}

© 中国科学院软件研究所 http://www.jos.org.cn

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

As Fig.6 shows, the prototype of VDHA_Grid has the following modules: (1) vdha module, which deals with the node membership management and virtual group tree maintenance; (2) user management module, which deals with the user management, user's password, user's public key and private key, user's access control, and so on; (3) service management module, which edits and views the service description and registers the services. (4) Create own key module, which creates the node's own public key and private keys. (5) Create a new Grid system module, which creates a new Grid system and this function is used only once by one node in a specific domain Grid system; and (6) event window, which monitors the messages and events of the Grid node. Figure 7 shows that after the client Web browser sends the requesting, the satisfied files distributed in all nodes in the VDHA_Grid system are listed in the browser.

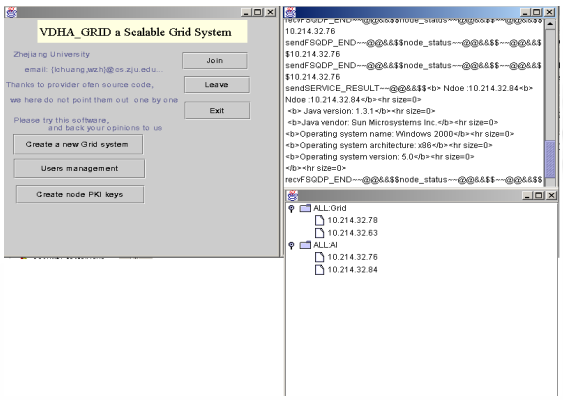


Fig.6 Main frame of VDHA_Grid

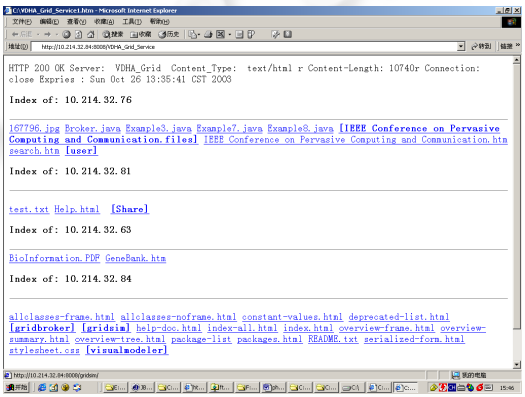


Fig.7 The result of file sharing service

2.7 Grid service description framework (GSDF)

In VDHA_Grid, the services can be dynamically appended into or deleted from the nodes. As infrastructures are different, MDS in Globus[5], and WSDL in Web Service are not enough to satisfy the needs. In VDHA_Grid, the services are used with three kinds of ways. (1) The service is simple, and the client end application can directly use it; (2) the client end application software uses the service's client-end API; (3) the client end program must be modified by programmer. So, the service description language must include service definitions which are understood by computer. This can be solved by the Ontology method. The language must include entities that are needed for implementing the service by SLMS. It also includes authorizations, accounting, protocol binding, and message format, etc. Therefore, GSDF must answer the following questions: (1) How does QDP use GSDF to find the services? (2) What protocol does the service bind and how does the protocol marshals? (3) How does SLMS use GSDF to implement the services? (4) How does the service account? (5) How does the service grant access right (authorization)? (6) What QoS does the service support? and so on.

We have designed an Ontology-based Grid Description language to solve the above problems. In OGSDL, the service is described by service profile, which describes what-is and how-to-do of the services. OGSDL is based on XML+RDF+OWL (see Fig.8). OGSDL describes services, operations, parameters and so on in ontology terms. From the services description, the computer can understand and deduce the facts with the aid of the background ontology and rules. Compared with DAML-S[13], OGDSL deals with the security and so on, and simplifies the description of service implementations.

© 中国科学院软件研究所 http://www.jos.org.cn

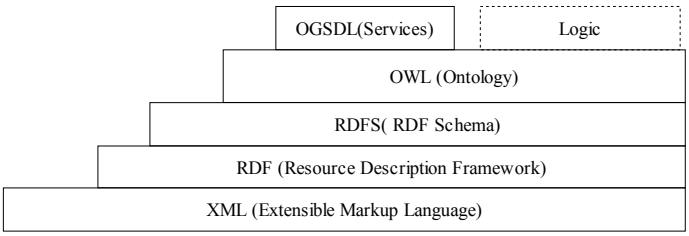
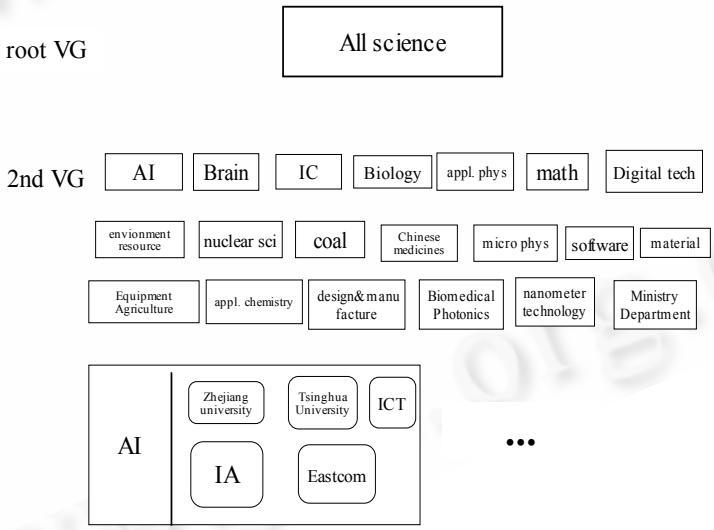


Fig.8 The layer of OGSDL

3 A Case Study

The virtual research projects granted by Ministry of Education of China are aimed to enhance the science and technology research by virtual cooperation via Internet. There are now 19 virtual organizations, each has a special domain. Each virtual organization has average 6 nodes which are located in Universities or research institutes. In order to combine these 19 organizations into an e-Science Grid system, we use VDHA to model this e-Science Grid system prototype (called as Chinese University e-Science Grid CUEG). Nineteen nodes are chosen from every 19 Vos, each plus one node located at Ministry of Education of China form an up-layer virtual group. Initially, the nodes located in the primary institutes of 19 VOs are chosen as gateway nodes (see Fig.9)



(Note: only nodes of AI are shown in the figure)

Fig.9 VDHA architecture of CUEG

In CUEG, we have implemented the prototype of file sharing service and knowledge retrieve service.

4 Conclusions

VDHA can solve the scale and autonomy problems. Some nodes can form a VO, and this VO can join the e-Science Grid without centralized administrator. In VDHA the messages are generally only concerned with the nodes of the three neighboring layer virtual groups, not with entire grid network. So, the e-Science Grid with VDHA has the possibility to become a huge net.

VDHA has high performance and exact discovery of resources and services. From the virtual group tree, we

© 中国科学院软件研究所 http://www.jos.org.cn

can know the detail information of every virtual group, so we can exactly and fully search the resources and services.

We have implemented a demonstration prototype VDHA_Grid. Our further work will focus on completing the VDHA_Grid and integrating it to CUEG, on enriching the services of CUEG, and on increasing the nodes of CUEG.

Acknowledgement This paper is supported by the Virtual Cooperative Research Project Granted by the Ministry of Education of China. Thanks specially colleagues and graduate students in our Laboratory for their discussion, cooperation, and contribution.

References:

[1] Taylor J. e-Science definition, 2002. http://www.e-science.clrc.ac.uk

[2] Foster I, Kesselman C, Tuecke S. The anatomy of the grid: Enabling scalable virtual organizations. International Journal of High Performance Computing Applications, 2001,15(3):200−222.

[3] Foster I, Kesselman C, Nick MJ, Tuecke S. The physiology of the grid: An open grid services architecture for distributed systems integration. 2002-02-17 http://www.globus.org/research/papers/ogsa.pdf

[4] Roure DD, Jennings N, Shadbolt N. Research agenda for the semantic grid: A future e-Science infrastructure, 2001. http://www.semantic grid.org/v1.9/semgrid.pdf

[5] Foster I, Kesselman C. Globus: A metacomputing infrastructure toolkit. International Journal of Supercomputer Applications, 1997,11(2):115−128.

[6] W3C, Web Services Architecture, 2003. http://www.w3.org/TR/2003/WD-ws-arch-20030808/

[7] Kant K, Iyer R, Tewari V. A framework for classifying peer-to-peer technologies. In: Proc. of the 2nd IEEE/ACM Int'l Symp. on Cluster Computing and the Grid (CCGRID02). 2002.

[8] Christensen E, Curbera F, Meredith G, Weerawarana S. Web Services Description Language (WSDL) 1.1 W3C. Note 15, 2001. http://www.w3.org/TR/wsdl

[9] Huang L, Wu Z, Pan Y. 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). Publishing House of Electronics Industry, 2002. 297−311.

[10] Message Sequence Charts (MSC), International Standard ITU-T Recommendation Z.120, 1992.

[11] Huang L, Wu Z, Pan Y. Virtual and dynamic hierarchical architecture: An overlay network topology for discovering grid services with high performance. Journal of Zhejiang University (Science), 2004,5(5):539−549.

[12] Larmouth J. ASN.1 Complete, 2003. http://www.oss.com/asn1/larmouth.html

[13] DAML-S, 2002. http://www.daml.org/services/

© 中国科学院软件研究所 http://www.jos.org.cn