CGSV*: An Adaptable Stream-Integrated Grid Monitoring System

Weimin Zheng[1*], Lin Liu[1†], Meizhi Hu[1†], Yongwei Wu[1*], Liangjie Li[1†], Feng He[1†], Jing Tie[2]

[1] Department of Computer Science and Technology, Tsinghua University
100084 Beijing, China
[*]{zwm-dcs, wuyw}@tsinghua.edu.cn
[†]{ll99, hmq02, liliangjie99, hefeng04}@mails.tsinghua.edu.cn
[2] Cluster and Grid Computing Lab, Huazhong University of Science and Technology,
430074 Wuhan, China
tiejing@gmail.com

Abstract. Grid monitoring is essential for the grid management and efficiency improvement. ChinaGrid Super Vision (CGSV) is proposed for ChinaGrid to collect status information of each entity (such as resources, services, users, jobs, Network), and provide corresponding information data query and mining services. In this paper, CGSV architecture and its components are discussed. CGSV is featured by data stream integration and adaptability to cope with dynamic measurement data and multiform query requirements. Measurement data can be accessed quickly and easily through WSRF-compliant services in CGSV. Transfer and control protocols are brought forward to facilitate data stream querying and runtime producer configuration in CGSV.

1. Introduction

1.1 ChinaGrid and CGSP

Grid computing has become the trend of distributed computing and Internet applications. As a problem solving mechanism, the grid supports geographically scattered communities to form Virtual Organizations [2], in order to achieve sharing and coordination of heterogeneous resources and to provide a virtual uniform application interface.

ChinaGrid (China Education and Research Grid) is the largest grid computing project in China, which is launched by the Ministry of Education (MoE) of China in 2002[1]. ChinaGrid aims to provide the nationwide grid computing platform and services for research and education purpose among 100 key universities in China. ChinaGrid Support Platform (CGSP) is developed for this ambitious goal. CGSP

* This work is supported by ChinaGrid project of Ministry of Education of China, Natural Science Foundation of China under Grant 60373004, 60373005, 90412006, 90412011, and Nation Key Basic Research Project of China under Grant 2004CB318000

provides a set of tools for ChinaGrid application developers and specific grid platform constructors. CGSP organizes grid resources in several domains. Each domain has a complete set of grid components to be able to function relative independently. ChinaGrid SuperVision (CGSV) is designed and developed based on CGSP and provides monitoring functions for ChinaGrid.

1.2 Monitor Requirements and CGSV

Grids are large-scale distributed system, featured by dynamic and complex, which require monitoring system running on to track status information of system resources (e.g. hardware, network, services) and further to perform analysis and optimization. Grid monitoring differs from traditional cluster monitoring mainly in that the former require scalable support for both pull and push data delivery model that may be distributed across organizations, together with extensibility and self-description support of data format for interoperability [3].

In ChinaGrid, monitoring system is an essential part to keep such a complex distributed system efficient. Most CGSP components, including Job Manager, Storage Manager and Information Center, require a monitoring system to provide system status information for different purposes. CGSV is then designed for monitoring ChinaGrid. In addition to common inventory tracking tasks, ChinaGrid also require CGSV to have scalable support to different types of data request, an efficient approach of data processing and transmission must be developed. Moreover, CGSV is also required to be able to dynamically change its monitoring behavior, that is to say, to change monitor entity metadata.

To cope with the above requirements, CGSV is designed to be an adaptable, stream-integrated grid monitor system. A transfer and control protocol is designed in order to efficiently transfer various types of measurement data and perform modification over producer behavior in a unified way. Re-publishers are stream oriented, in that they are designed to support predicate-based processing over measurement data streams and SQL-like stream query interfaces.

1.3 Roadmap

In Section 2, CGSV's requirements, objective and position in ChinaGrid are overviewed. In Section 3, basic system architecture is given, followed by the detail design of system building blocks. Section 4 introduce the stream integration attempts in CGSV. Finally, Section 5 compares some related works correlated with our design and implementation, and in Section 6 a conclusion is summarized and future plan is listed.

2. Overview

2.1 Requirements and Objective

The final goal of CGSV is to implement a monitor and management module in ChinaGrid, which enable users or other grid modules in ChinaGrid to perform different level of system performance monitoring, analysis and optimization in a flexible way. Hardware resources, network condition, grid services and job status, the four main targets of ChinaGrid will be monitored in CGSV.

2.2 CGSV vs CGSP

CGSV will be implemented to be a control tower of ChinaGrid. The function of CGSV is distinguished from Information Center, but they also rely on each other. CGSV collects measurement data from hardware, Service Container, Job Manager, Storage Manager, and provides these data to Information Center. Domain monitoring services of CGSV acts like other system services in ChinaGrid, which rely on Information Center's domain topology information to locate monitoring services deployed in other domains. CGSV puts emphasis on dynamic monitor information while Information Center focuses relatively static information.

3. Architecture

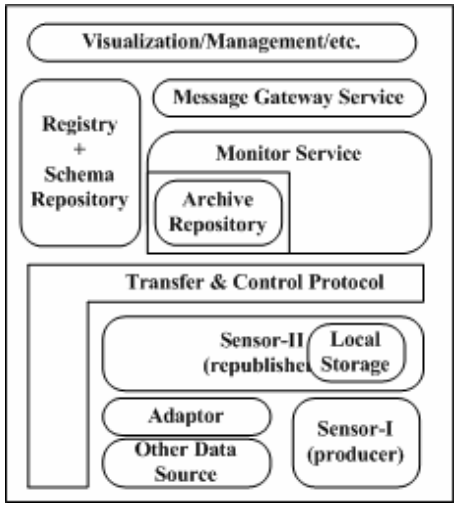

Fig. 1. Basic Architecture of CGSV Module Deployment

CGSV is designed based on the Grid Monitoring Architecture [4] proposed by the Global Grid Forum. As Fig.1 shown above, CGSV is built up by several components,

which can be divided into 3 layers, that is, collection layer, service layer and presentation layer. On the collection layer, sensors are deployed for measurement data collection. Sensor-I is responsible for collecting hardware resource information. Other sources of information including network condition and dynamic service information or data generated by other monitor tools are wrapped by an adaptor to unify the data format. Above the data interface based on a transfer and control protocol, lies the service layer. In each domain of ChinaGrid, this layer is presented as a logical domain monitor center, where Domain Registry Service and Monitor Services are deployed. A Message Gateway Service is provided to alleviate the communication cost of monitor services since over numbered notification and subscription will disastrously decease service performance. The top layer is the presentation layer, where data analysis, visualization work and management can be performed. Detailed description of sensors, protocol and stream-integration design will be discussed in the next section.

3.1 Collection Layer

For the purpose of both compactness and runtime controllability, we develop our own contributed sensors for measurement data collection.

Unlike many existing monitor tools, the most significant characteristic of our sensor is runtime configurable, which means that the monitor metadata, such as each metric's switch, collection frequency and granularity, is able to be changed over runtime on demand. For example, we demand turning off all the resource monitoring metrics except CPU load and also lower down the information collecting frequency to alleviate intrusiveness on some machines with heavy load. For many other monitor tools, configuration is written in files and is load only at startup; therefore the required action needs us to login on that computing node, shutdown the tool, change configuration file and start the tool again. This complicated work is not flexible for the dynamic environment of the grid, where similar scenarios are envisioned to occur frequently. In contrast, in CGSV, This action only needs us or grid system components to send a command according to our protocol. The sensors will then automatically change their configuration.

Configuration file is also used in our implementation, for initialization and logging configuration when changes occur. In other words, this file is the real-time hard-disk backup of sensor configuration, and is read only at startup. The configuration does not exist in any other materialized form even in memory.

There are 2 main types of sensors called sensor-I and sensor-II in CGSV. The difference between the 2 types is their function and deployment location in resource and network monitoring. For corresponding components in GMA, sensor-I is the actually producer and sensor-II can be treated as re-publisher.

Sensor-I is deployed on computing nodes of clusters and any PC resources. They are responsible for collecting resource information. Broadcast discovery and data transmission between Sensor-I are performed via UDP packets with the format specified in a message protocol.

Sensor-II is deployed on front-end node of clusters and any PC resources. They are responsible for pulling resource information from sensor-I like data sources (sensor-I

also supports the pushing manner) through UDP messages and processing incoming data request through TCP massages. Dynamic information of web services are available through APIs of Service Container, so we can treat the APIs as sensors and wrap them with an adaptor, so that measurement data can go through a unified path. In addition, Sensor-II is also responsible for processing control messages and adjusting behaviors of Sensor-I. Sensor-II can also collect information from other sensor-like components. All the messages are under the protocol that will be discussed in 3.2.

Sensor-II can be connected hierarchically, but generally we do not advocate this method in CGSV for resources are autonomous and independent in ChinaGrid. So within each domain of ChinaGrid, Sensor-II are deployed flatly and connected to a single domain monitoring center.

3.2 Transfer and Control Protocol

A message protocol is designed for both measurement data transmission and sensor control. Inspired by the classical File Transfer Protocol (FTP) [14] and Supermon project [9], our protocol is a client-server protocol, based on symbolic expressions (or s-expressions). The underlying transfer protocol varies from UDP to TCP as described in the above paragraph.

S-expressions originated from LISP as a recursively defined, simple format for data representation. This format of data has the following features

- Extensibility: the protocol can be extended for new data types and new type of commands, which allows the system to be capable to evolve.
- Self-descriptive: each measurement data is associated with its metric name and timestamp, so the packet can be independently interpreted without any other knowledge, therefore increases the system's interoperability.
- Compactness: Though the packets are self-descriptive, the format is very compact comparing with XML. This feature saves network transmission bandwidth and memory cost for protocol interpretation, thus decrease the intrusiveness to host systems.
- Architecture independence: This is achieved by plain textual representation, which facilitates system portability.

For the convenience of protocol parsing, not only data and schema packets, but also command packets, including query, control and result, are encoded in s-expressions. So within each domain of ChinaGrid, all monitoring messages transferred are packets in the form of s-expression protocol. This unified data transmission and control method simplifies the implementation of monitoring components and naturally makes protocol interpretation and protocol execution logically separated.

This effort also makes CGSV components loose coupled and easy to collaborate with other monitor systems. Table 1 lists the basic packet types:

Table 1. Five basic packets types implemented in CGSV protocol

Type	Packet example	Purpose	Comments
QUERY	(QUERY (get 1.2.3.4	Issue a data query	If this IP is a cluster, all

	*))	request for all measurement data of host 1.2.3.4	back-end nodes' information are returned
DATA	(DATA (hostname 1.2.3.4)(timestamp 11 15715626)(OSType 1 10)(CPUNumber 2 10))	Data packet indica-ting 2 monitor items (OS type and CPU number) with host IP and collect time.	metric info tuple is composed by metric name, value and time difference with the complete timestamp.
SCHEMA	(SCHEMA (hostname 1.2.3.4)(OSType 1 1500)(CPULoad 0 15))	Schema packet indic-ating 2 metrics are supported on host 1.2.3.4 with their switches and monitor intervals	Metric schema tuple is composed by the name, switch flag and monitor interval in seconds
CTRL	(CTRL (close 1.2.3.4 4))	Issue a control request to switch MemFree metric off on host 1.2.3.4	"4" is predefined number for MemFree metric, full metric name is also accepted.
RESULT	(RESULT (1 (get 1.2.3.4 *)))	indicates execution result (1) of command "get 1.2.3.4 *"	"1" is predefined error code for success

3.3 Service Layer

3.3.1 Registry

In each domain of ChinaGrid, we have a logical domain monitor center, where registry, archive module and monitor services are deployed.

Registry of CGSV performs 2 tasks. One is for producer/re-publisher registration, and to provide a lookup service for consumers to locate; the other task is to store producers' metric schema for system extension and modification. Since the adaptable implementation of sensors and protocol allows producers to accept control packets from any trusted source, the schema held by Registry needs to be synchronized periodically.

3.3.2 Archive

Archiving is an optional component. It periodically acquires measurement data from all data sources and stores them as historical information in DBMS. This mechanism works similar as registry schema synchronization. The different is that archive is much more costly and storage size increases quickly, so a distributed DBMS is used to share the loads when the domain grows larger.

3.3.3 Services and Message Gateway

Domain Monitor Services are WSRF-compliant services which are responsible for providing monitor information and management information interfaces. WSDM specification [6] has been studied and applied on monitor services for management issues. Each domain is treated as a WSDM manageability capability.

To alleviate the communication cost of monitor services, a message gateway service is used only for transmitting request and response between monitor services and grid users. As a result, data processing and data transmission are separated, and then service load is distributed.

3.4 Presentation Layer

Visualization work is implemented by http server plus servlet and java applet, to perform several forms (tables, histograms) of data presentation. Measurement data are retrieved from monitor services or monitoring service gateway. Users can view the grid by simply browsing the web pages. GIS (Geographical Information System) is introduced to locate resources from their geographical location on maps. An open source toolkit JFreeChart [16] is used for diagram plotting support. Basic Real-time visualization is implemented for dynamic resource information. To reveal relationship between metrics, diagrams correlated with 2 or more metrics are designed for intuitive data analysis. Management actions can also be performed through the GUI client.

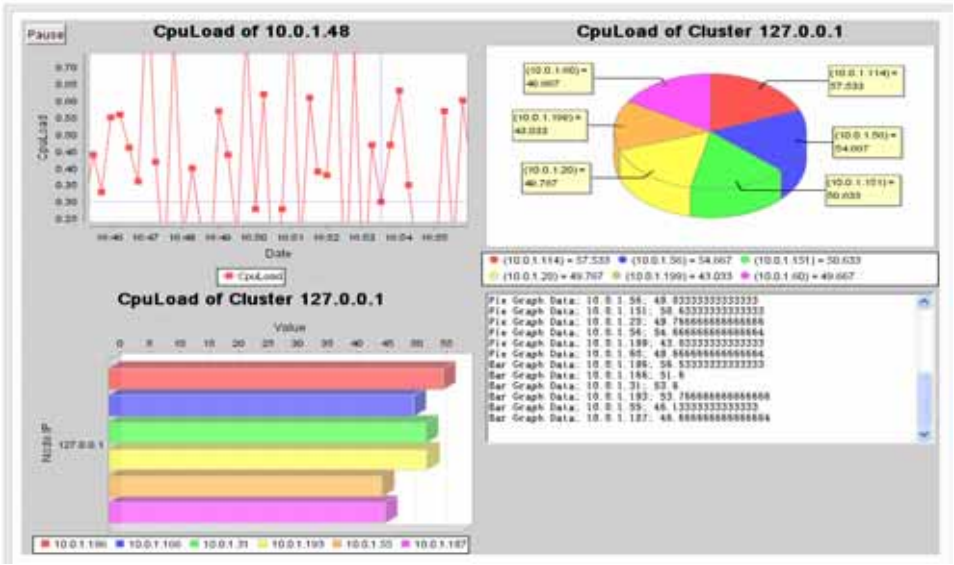

Fig. 2. Various visualization forms of CGSV implementation

4. Stream-orientated Scheme

Data Stream systems have been proven to be suitable for monitoring applications [12]. Research on data stream management system design and stream query model have attracted great effort of work and mature formal theory is proposed [11]. In CGSV, system monitor information is treated as a huge stream, which is composed of several levels of data streams. CGSV focuses on the stream-like features of monitor information and behaviors, such as trigger-oriented, real-time requirement. Two

extreme viewpoints are avoided here. One is only to see the grid's instant status information, where instant data is not enough in many system usage scenarios such as failure analysis. The other is to view the grid as a virtual database, which often requires large storage and schema mapping and translation. This approach is feasible but often suffers from redundancy storage.

Stream integration in grid monitoring is a compromise of data storage, efficiency and functional capability. Stream processing should be put close to data source to distribute load and improve efficiency. For the sake of integration of any kinds of sensors, CGSV implements data stream on sensor-II, the actually re-publisher.

Fig. 3 shows the stream integration structure of re-publishers. Data streams come from Measure Data Puller, which pulls monitor information from producers. Data Stream Queue Manager holds two types of queues. Recent measurement information is kept in memory as buffer window queues, while outdated information is materialized in local storage. Since more recent information is usually more important and more frequently used, this division is reasonable. Both stream queues in memory and local storage form the input stream for processing. Queries coming from protocol interpreter give the system two input information. Data processing predicates are processed by Filter Inserter, and then are inserted to Predicate Manager. Predicate Manager maintains a predicate queue and also a processing plan by combination and optimization of predicates. Connection information is kept as channel queues. After processing of input streams, responses are sent to corresponding channel.

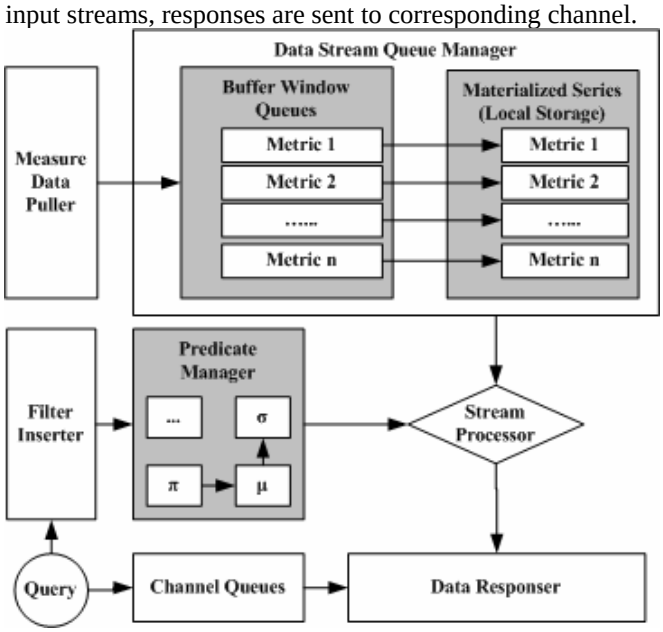

Fig. 3. Re-publisher Stream Integration Design

This structure has to coordinate with the proposed message passing protocol, in order to facilitate query parsing and data response transmission. The underlying Measure Data puller is a flexible component, which can be modified to combine any other monitor tools.

Stream query language is a set of SQL-like queries. The support to the language depends on the implementation of two components of Predicate Manager and Stream Processing. Single stream processing is supported currently.

5. Related Work

Grid monitoring is not a new issue. There are a large number of mature projects working on this area, but very few of them have a focus on stream integration for high efficiency and adaptability is seldom considered either. CGSV benefits from their efforts, and tries to naturally combine the outstanding features of some monitoring projects while avoiding their shortcomings. CGSV is made up as an adaptable and efficient grid monitoring framework based on ChinaGrid.

CGSV's design uses several monitoring tools and data stream projects as good references, integrates their features in one framework and also developed its own features. The feature of sensors' scalable broadcast discovery at cluster level is learned from Ganglia project [8]. Message protocol design is enlightened by Supermon [9]'s kernel mode protocol in the form of s-expression. Stream processing design is inspired by Aurora [12], a data stream project. Finally, MonALISA [10]'s intuitive and impressive visualization work, which has a rich set of visualization forms, has a great impact on CGSV presentation works. Besides, CGSV attempts to integrate data stream in re-publishers and has adaptable design of runtime configured sensors and extensible protocols for both data transfer and control.

The first two cluster monitor tools mentioned above has their problems in grid monitoring context. Ganglia has a registry-free arbitrary architecture with filter-free aggregation, so it can only be used as basic sensors. However, Ganglia sensors are not runtime configurable. Supermon uses a statically configured hierarchy of point-to-point connections which makes it less scalable.

Relational Grid Monitoring Architecture (R-GMA) is a grid monitor system considered stream integration problems. R-GMA perceives Grid monitoring as a data integration problem, and extends GMA by choosing the relational data model. They have performed some research on stream integration and developed basic techniques. [13] However, they ignore the activity of monitor information, and treat the data statically as a virtual database thus do not benefit from stream adequately.

6. Conclusion

CGSV is a complete set of grid monitor solution for ChinaGrid. In this paper, we first introduce the basic CGSV architecture, along with some detail design and implementation issues on system building blocks. CGSV focused on sensor controllability, adaptable message protocol and stream integration on re-publishers, and proposed a flexible mechanism for grid monitoring.

Our future plan of CGSV considers 4 research points.

- Data analysis, which assists decision making, and hence makes the behavior of our adaptable sensors automatic

- Scalable optimization of data stream model to cope with large number of queries and predicates.
- Security is also an important issue to be considered. The message protocol needs security data transfer to restrict access to sensors and re-publishers, and to protect sensitive measurement data.
- Measurement data precision representation and synchronization.

References

1. Hai, J.: ChinaGrid: Making grid computing a reality. Digital Libraries: International Collaboration and Cross-Fertilization, Proceedings, Vol. 3334. Springer-Verlag Berlin (2004) 13-24
2. Foster, I., Kesselman, C., et al.: The anatomy of the grid: Enabling scalable virtual organizations. International Journal of High Performance Computing Applications, 15(3). Sage Publications Inc (2001) 200-222
3. Zanikolas, S. and Sakellariou, R.: A taxonomy of grid monitoring systems. Future Generation Computer Systems, 21(1). Elsevier Science Bv (2005) 163-188
4. Tierney, B., Aydt, R., et al: A Grid Monitoring Architecture. GWDPerf-16–3, Global Grid Forum, August 2002. http://wwwdidc.lbl.gov/GGF-PERF/GMA-WG/papers/GWD-GP-16-3.pdf
5. Web Service Resource Framework (WSRF): http://www.globus.org/wsrf/
6. OASIS Web Services Distributed Management (WSDM): http://www.oasis-open.org /committees/tc_home.php?wg_abbrev=wsdm
7. GT Information Services: Monitoring & Discovery System (MDS): http://www.globus.org /toolkit/mds/
8. Massie, M. L., Chun, B. N., et al.: The ganglia distributed monitoring system: design, implementation, and experience. Parallel Computing, 30(7). Elsevier Science Bv (2004) 817-840
9. Sottile, M. J. and Minnich, R. G.: Supermon: A High-Speed Cluster Monitoring System. Proceedings of the IEEE International Conference on Cluster Computing. Washinton D.C. (2002) 39 IEEE Computer Society
10. Newman, H. B., Legrand, I. C., et al.: MonALISA: A Distributed Monitoring Service Architecture. Computing in High Energy and Nuclear Physics (CHEP03). La Jolla, California. (2003)
11. Babcock, B., Babu, S., et al.: Models and issues in data stream systems. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Madison, Wisconsin. (2002) 1-16 ACM Press
12. Carney, D., Cetintemel, U., et al.: Monitoring Streams - A New Class of Data Management Applications. Proceedings of Very Large Databases (VLDB). HongKong. (2002)
13. Cooke, A., Gray, A. J. G., et al.: Stream integration techniques for Grid monitoring. Journal on Data Semantics Ii, Vol. 3360. Springer-Verlag Berlin (2005) 136-175
14. Postel, J., Reynolds, J.: File Transfer Protocol (FTP). Available from http://www.ietf.org /rfc/rfc959.txt
15. JFreeChart Project: http://www.jfree.org/jfreechart/index.html
16. ChinaGrid Project: http://www.chinagrid.edu.cn