A POINT-TO-POINT BASED CLOUD WORKFLOW SYSTEM

A.M.K.KANNA BABU
Dept. of IT
Sir C.R.Reddy college of Engg,Eluru, India
Email: kumarkanna1@gmail.com

N.PRASAD
Dept. of IT
Sir C.R.Reddy college of Engg,Eluru, India
Email: nprasad9999@gmail.com

K.LAKSHMAJI
Dept. of IT
Sir C.R.Reddy college of Engg,Eluru, India
Email: kotlalakshmaji@gmail.com

ABSTRACT
Workflow systems are designed to support the process automation of large scale business and scientific applications. In recent years, many workflow systems have been deployed on high performance computing infrastructures such as cluster, point to point , and grid computing. One of the driving forces is the increasing demand of large scale instance and data/computation intensive workflow applications (large scale workflow applications for short) which are common in both e-Business and e-Science application areas. The data and computation intensive pulsar searching process in Astrophysics. Generally speaking, instance intensive applications are those processes which need to be executed for a large number of times sequentially within a very short period or concurrently with a large number of instances. Therefore, large scale workflow applications normally require the support of high performance computing infrastructures (e.g. advanced CPU units, large memory space and high speed network), especially when workflow activities are of data and computation intensive themselves. In the real world, to accommodate such a request, expensive computing infrastructures including such as supercomputers and data servers are bought, installed, integrated and maintained with huge cost by system users.

1. INTRODUCTION

Workflow systems are designed to support the process automation of large scale business and scientific applications. In recent years, many workflow systems have been deployed on high performance computing infrastructures such as cluster, point-to-point and grid computing. One of the driving forces is the increasing demand of large scale instance and data/computation intensive workflow applications (large scale workflow applications for short) which are common in both e-Business and e-Science application areas.

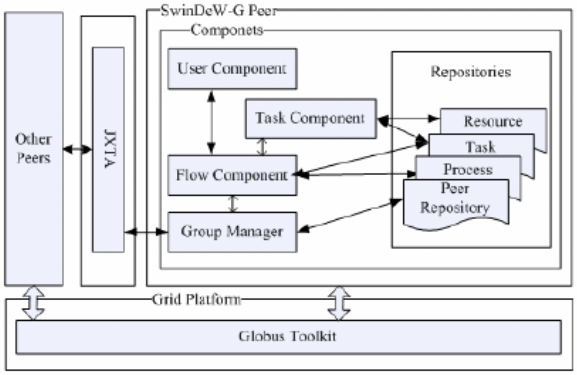
Besides scalable resources, another principal issue for large scale workflow applications is decentralized management. In order to achieve successful execution, effective coordination of system participants (e.g. service providers, service consumers and service brokers) is required for many management tasks such as resource management (load management, workflow scheduling), QoS (Quality of Service) management, data management, security management and others. One of the conventional ways to solve the coordination problem is centralized management where coordination services are set up on a centralized machine. All the communications such as data and control messages are transmitted only between the central node and other resource nodes but not among them. However, centralized management depends heavily on the central node and thus can easily result in the performance bottleneck. Some others common disadvantages also include: single point of failure, lack of scalability and the advanced computation power required for the coordination services. To overcome the problems of centralized management, decentralized management where the centralized data repository and control engine are abandoned, and both data and control messages are transmitted between all the nodes through general broadcast or limited broadcast communication mechanisms. Thus the performance bottlenecks are likely eliminated and the system scalability can be greatly enhanced point to point is a typical decentralized architecture. However, without any centralized coordination, pure point to point (unstructured decentralized) where all the point nodes are communicating with each other through complete broadcasting suffers from low efficiency and high network load. Evidently, neither centralized nor unstructured decentralized management is suitable for managing large scale workflow applications since massive communication and coordination services are required. Therefore, in practice,
structured point to point architecture is often applied where a super node acts as the coordinator peers for a group of points. Through those super nodes which maintain all the necessary information about the neighboring nodes, workflow management tasks can be effectively executed where data and control messages are transmitted in a limited broadcasting manner. Therefore, structured decentralized management is more effectively than other for managing workflow applications.

2. LARGE SCALE WORKFLOW APPLICATIONS

Here, we present two examples, one is from the business application area (insurance claim) and the other one is from the scientific application area (pulsar searching).Insurance claim: Insurance claim process is a common business workflow which provides services for processes of such as insurance under employee benefits including, for example, medical expenses, pension, and unemployment benefits.

Due to the distributed geographic locations of a large number of applicants, the insurance offices are usually deployed at many locations across a wide area serving for a vast population. Despite the differences among specific applications, the following requirements are often seen in large/medium sized insurance companies: supporting a large number of processes invoked from anywhere securely on the Internet, the privacy of applicants and confidential data must be protected avoiding management of the system at many different locations due to the high cost for the setting and ongoing maintenance; being able to serve for a vast population involving processes at the minimum scale of tens of thousands per day, i.e. instance intensive; and for better quality of service, needing to handle workflow exceptions appropriately, particularly in the case of instance-intensive processes.

3. ARCHITECTURE OF SWINDEW-C POINTS



The Group Manager is the interface between the peer and JXTA. In JXTA, all communications are conducted in terms of peer group, and the Group Manager maintains the peer groups the peer has joined. The information of the peer groups and the peers in them is stored in the Peer Repository. While a SwinDeW-G point is implemented as a grid service, all direct communications between points are conducted via point to point. Points communicate to distribute information of their current state and messages for process control such as heartbeat, process distribution, process enactment etc.

The User component is the interface between the corresponding workflow users and the workflow environment. In SwinDeW-G, its primary function is to allow users to interfere with the workflow instances when exceptions occur.

Globus Toolkit serves as the grid service container of SwinDeW-G. Not only a SwinDeW-G point itself is a grid service located inside Globus Toolkit, the capabilities which are needed to execute certain tasks are also in forms of grid services that the system can access. That means when a task is assigned to a peer, Globus Toolkit will be used to provide the required capability as grid service for that task.

Data management component in SwinDeW-C consists of three basic tasks: data storage, data placement and data replication. *Data Storage*: In this component, a dependency based cost-effective data storage strategy is facilitated to store the application data . The strategy utilizes the data provenance information of the workflow instances. Data provenance in workflows is a kind of important metadata, in which the dependencies between datasets are recorded .The dependency depicts the derivation relationship between the application datasets.

In cloud workflow systems, after the execution of tasks, some intermediate datasets may be deleted to save the storage cost, but sometimes they have to be regenerated for either reuse or reanalysis. Data provenance records the information of how the datasets have been generated. Furthermore, regeneration of the intermediate datasets from the input data may be very time consuming, and therefore carry a high computation cost. With data provenance information, the regeneration of the demanding dataset may start from some stored intermediated datasets instead. In a cloud workflow system, data provenance is recorded during workflow execution.

Taking advantage of data provenance, we can build an Intermediate data Dependency Graph (IDG) based on data provenance. All the intermediate datasets once generated in the system, whether stored or deleted, their references are recorded in the IDG. Based on the IDG, we can calculate the generation cost of every dataset in the cloud workflows. By comparing the generation cost and storage cost, the storage strategy can automatically decide whether a dataset should be stored or deleted in the cloud system to reduce the system cost, no matter this dataset is a new dataset, regenerated dataset or stored dataset in the system.

4. SECURITY MANAGEMENT IN SWINDEW-C

To address the security issues for the safe running of SwinDeW-C, the security management component is designed. As a type of typical distributed computing system, trust management for SwinDeW-C points is very important and plays the most important role in security management. Besides, there are some other security issues that we should consider from such as user and data perspective. Specifically, there are three modules in the security management component: trust management, user management and encryption management system.

Trust management: The goal of the trust management module is to manage the relations between one SwinDeW-C point and its neighbouring points. For example, to process a workflow instance, a SwinDeW-C point must cooperate with its neighbouring points to run this instance.

User management: the user management module is an essential piece in every system. In SwinDeW-C, a user base is a database which stores all user identity and log information that submit service requests. In addition, an authority manager controls the permissions for users to submit some special service requests.

Encryption management System: Given SwinDeW-C points are located within different geographical local networks, it is important to ensure the data security in the process of data transfer by encryption. In SwinDeW-C, we choose the PGP tool GnuPG to ensure secure commutation.

5. CONCLUSIONS AND FEATURE WORK

Large scale sophisticated workflow applications are commonly seen in both e-Business and e-Science areas. Workflow systems built on high performance computing infrastructures such as cluster, point to point and grid computing are often applied to support the process automation of large scale workflow applications. However, two fundamental requirements including scalable resources and decentralized management have not been well addressed so far. Recently, with the emergence of cloud computing which is a novel computing paradigm that can provide virtually unlimited, easy-scale computing resources, cloud workflow system is a promising new solution and thus deserves systematic investigation.

6. REFERENCES

1. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A.
Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report, University of California at Berkeley, 2009.

2. R. Bose and J. Frew, "Lineage Retrieval for Scientific Data Processing: A Survey," *ACM Comput. Surv.*, vol. 37, no. 1, pp. 1-28, 2005.

3. R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud Computing and
Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599-616, 2009.

4. J. Chen and Y. Yang, "A Taxonomy of Grid Workflow Verification and Validation," *Concurrency and Computation: Practice and Experience*, vol. 20, no. 4, pp. 347-360, 2008.

Mr.A.M.K.KANNA BABU is working as an Asst.Professor, in IT Department, Sir C. R.Reddy College of Engg, Eluru, A.P., and India. He has received his B.Tech(CSE) from Sir C R Reddy College of Engineering, Eluru, and M.Tech. (SE) from Avanthi Institute of Engineering and Technological. Makavarapupalem.A.P., INDIA. His research interests include Image Processing, Data Mining, Networks Security, Web security, Software Engineering, Computer Networks and Wireless Networks.

Mr. N. Prasad is working as an Asst. Professor, in IT Department, Sir C. R. Reddy College of Engg, Eluru, A.P., and India. He has received his B.Tech (CSE) from SVH College of Engineering, Machilipatnam and M.Tech (CSE) from Jawaharlal Nehru Technological College of Engg.Hyderabad.(JNT University),Hyderabad, A.P., INDIA. His research interests include Data Mining, Networks security, Web security, WSNs and Computer Networks.

Mr. Lakshmaji Kotla is working as an Asst. Professor, in IT Department, Sir C. R. Reddy College of Engg, Eluru, A.P., India. He has received his M.Sc(CS) from Gowri P.G College Visakhapatnam and M.Tech (CSE) from Vasavi

Engineering college Tadepalligudem (JNTUK University), Kakinada, A.P., INDIA. His research interests include Image Processing, Data Mining, CryptoGraphy & Network security, Biometrics, Cloud Computing and Software Engineering..