A Stochastic Control Model for Hierarchical Grid Service*

Zhimin Tian, Li Liu, Yang Yang, and Zhengli Zhai

School of Information Engineering,
University of Science and Technology Beijing, Beijing, China
t_zhm@163.com

Abstract. In this paper, we introduce a model for deployment and hosting of a hierarchical gird service wherein the service provider must pay to a resource provider for the use of resources. Our model produces policies that balance the number of required resources with the desire to keep the cost of hosting the service to a minimum. In each layer of our framework, we quantify the cost increase of reserved resources caused by the fluctuation of the users' demand. A stochastic control algorithm is cast in order to resolve the problem. The results show that the model makes good decisions in the face of such uncertainties as random demand for the service.

1 Introduction

The success of OGSA (Open Grid Service Architecture) and web services has influenced grid applications [1]. Grid application designers are now beginning to make use of software services that provide a specific functionality to the application, such as solving a system of equations or performing a simulation remotely. Grid applications that make use of such services require consistent response time and high availability of those services. The service provider, who develops the service and its interface, may charge users through subscriptions to the service [2]. In turn, we assume that there is a cost to the service provider for maintaining the presence of a service in the grid. This cost is charged to the service provider by the owner and maintainer of the computational resources, the resource provider[3]. If there were no costs to maintain the presence of a grid service, then the service provider could simply deploy the service in as many places as possible and leave it running. Therefore, the service provider must balance the demand for service with the desire to keep the cost of providing it to a minimum. This work focuses on controlling the cost.

The amount of resources needed may vary over time and is a function of the demand for the service and the compute intensive nature of the service. We address the situation where the service demand and the execution time to process the service requests are unknown, but can be estimated. Even though the service provider will know the processing requirements for a typical invocation of the service, the execution time of any particular instantiation of the service can vary due to input data

* This work has been supported by National Natural Science Foundation of China. (No. 90412012).

H. Jin, D. Reed, and W. Jiang (Eds.): NPC 2005, LNCS 3779, pp. 72 – 79, 2005.

dependencies as well as resource contention with other services if, as is likely in a grid, the service is deployed in a time-sharing environment.

In this paper, we propose a layered model for service grid. The models are designed for a service grid focusing on business intelligence services that often involve a lot of data and many complex algorithms. The service Grid provides an economic platform for business intelligence services. In the model, the number of resources each tier provides is larger than that the users demand, because their demand is uncertain. i.e. the demand variability increases when it moves up a chain. This will make the cost increase. Our work is to control the cost by using the stochastic control theory.

The paper is organized as follows. Related works are reviewed in Section 2. In Section 3, a resource schedule framework of service grid is described. In Section 4, the stochastic control algorithms are discussed in details. Some simulation experiments are presented in Section 5. Finally, we draw some conclusions in Section 6.

2 Related Works

A number of works have proposed service-oriented architectures and have tested high-performance applications in those environments [4][5]. Weissman and Lee presented an architecture and middleware for dynamic replica selection and creation in response to service demand [6]. Their work answers the questions of when and where to deploy a grid service. In contrast, this work focuses on the question of how many resources are required to host a grid service in the presence of random demand and execution times. Buyya et. al.[7] and Wolski et. al. [8] examined the use of supply- and demand-based economic models for the purpose of pricing and allocating resources to the consumers of grid services. In this article we assume a supply- and demand-based economy in which both software services and computational resources are in demand. In particular, we assume a separation of interests between the service provider and the resource provider. The service provider obtains the necessary computational resources at a cost. The user then, is only concerned with the software services that are required for the application, rather than negotiating directly with a resource owner for computing time.

3 Grid Service Hierarchical Framework

Fig. 1 shows a layered architecture of a service Grid. Logically, the architecture is divided into four tiers: the User Tier, the Grid Tier, the Admin Domain Tier and the Node Tier. The Node Tier can be a computer, a service provider and a storage resource. It provides all kinds of resources for the upper tier. The Admin Domain Tier consists of machine groups, named as Admin Domains (AD), in which all nodes belong to one organization. For example in Fig. 1, AD1 belongs to the Computer Center and AD2 belongs to the Department of Computer Science. On the one hand, each AD can be regarded as a whole system, and all nodes in it have a common objective. On the other hand, an AD can fully centrally control the resources of its nodes but cannot operate the resources of nodes in the other ADs directly. In this view, all nodes are cooperative in the same AD. The ADs Tier not only provides service for the Grid Tier, but also reserves resource of the Node Tier. A Grid Service Tier can have many

ADs connected together and have good collaboration and trust relationship between the ADs. For example, Grid Service1 in Fig. 1 can be a Grid in Hong Kong and Grid Service2 is a Grid in Beijing, China. However, Grid services are independent from each other, the user can submit tasks to a Grid from its Portal. Likewise, the Grid Service Tier provides service for users.

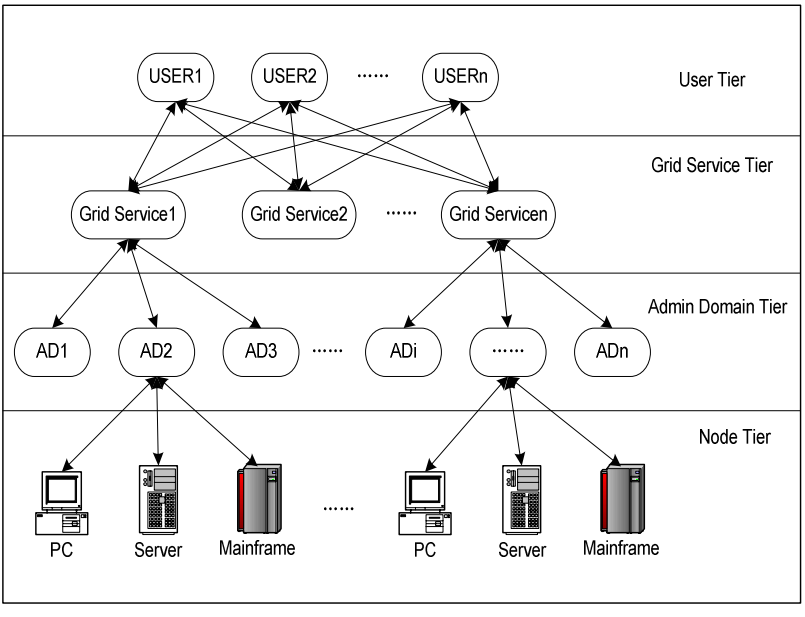

Fig. 1. The hierarchical framework for service grid

In this framework, the User Tier will subscribe services that the Grid Service Tier provides. Generally, what the Grid Service Tier provides is more than what the User Tier demands, or the application request can't be executed in time once some services halt. We define the offset as surplus resources. So is the Admin Domain Tier. Therefore, the surplus resources will become greater and greater from User Tier to Node Tier when the user's requests increase. The cost is also higher and higher. Our work is study how to control the cost by stochastic control theory.

4 Modeling for Hierarchical Grid Service

4.1 Modeling for Grid Service Chain

At first, we define the variables as follows:

$u^s_{1,k}$ ——denotes quantity of user demand; n-dimensional vector;
$x^s_{1,k}$ ——denotes surplus resources in grid service tier, n-dimensional vector;
$d_{1,k}$ ——denotes certainty of user demand, n-dimensional vector;
$u^s_{2,k}$ ——quantity of services that the Grid Service Tier subscribes in Admin

Domain Tier, m-dimensional vector;

$x_{2,k}^{s}$ ——the surplus resources of the Admin Domain Tier, m-dimensional vector;

L——m×n matrix;

$Lu_{1,k}^{s}$ ——quantity of user demand in the Admin Domain Tier.

So our system model is:

$$x_{1,k+1}^{s} = x_{1,k}^{s} + u_{1,k}^{s} - d_{1,k} \tag{1}$$

$$x_{2,k+1}^{s} = x_{2,k}^{s} + u_{2,k}^{s} - Lu_{1,k}^{s} \tag{2}$$

where $L=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ l_{21} & l_{22} & \cdots & l_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ l_{m1} & m2 & \cdots & l_{mn} \end{bmatrix}$, $l_{ij} \geq 0, i = 1,2,\cdots,m; j = 1,2,\cdots,n. \sum_{i}^{m} l_{ij} = 1$

L's row vector ($l_{i1}, l_{i2}, \cdots, l_{in}$) is a weight vector. It represents proportion of n users request in i node of the Admin Domain Tier.

Equation (1) denotes dynamic process in Grid Service Tier. Equation (2) denotes dynamic process in Admin Domain Tier. They are presented by matrix formal:

$$x_{k+1}^{s} = x_{k}^{s} + Bu_{k}^{s} + d_{k} \tag{3}$$

where $B = \begin{bmatrix} I & 0 \\ -L & I \end{bmatrix}$, $d_k = \begin{bmatrix} -d_{1,k} \\ 0 \end{bmatrix}$

When the user's requests change, w_k, the variety moves down from user tier to node tier, further, it will be greater. So, the state equation may be presented as:

$$x_{k+1}^{f} = x_{k}^{s} + Bu_{k}^{f} + d_{k} + w_{k} \tag{4}$$

where d_k is part of the certainty of the users' demand, n-vector, w_k is part of the uncertainty of users' demand, n-vector. Because of the change of the users' demand, the state variable, x_k^s, and the control variable, u_k^s produce variety. They turn into x_k^f, u_k^f respectively.

Now, we define the offset as follows:

$$x_k = x_k^f - x_k^s \tag{5}$$

$$u_k = u_k^f - u_k^s \tag{6}$$

Here, the system offset equation is:

$$x_{k+1} = x_k + Bu_k + w_k \tag{7}$$

where w_k is Gauss white noise, i.e. $w_k \sim N(0, \sigma_w^2)$, its covariance matrix is R_1.

In this hierarchical grid service system, the surplus resources offset of each tier is obtained by observation. So, it has noise effect:

$$y_k = x_k + v_k \tag{8}$$

where y_k is an observation value, (n+m) dimensional vector; v_k is white noise, (n+m) dimensional vector, i.e. $v_k \sim N(0, \sigma_v^2)$, its covariance matrix is R_2.

4.2 Quantifying for Layered Grid System

The offset is used to quantificationally describe the situation which users' demand change affect each tier of hierarchical grid service system. Furthermore, the offset will be amplified from top to bottom. This is presented as follows:

$$y_1^2 = \frac{x_{1,k}^T Q_1 x_{1,k} + u_{1,k}^T u_{1,k}}{w_k^T w_k} \tag{9}$$

$$y_2^2 = \frac{x_{2,k}^T Q_2 x_{2,k} + u_{2,k}^T u_{2,k}}{w_k^T w_k} \tag{10}$$

where, Q is not negative definite matrix, and Q=diag(Q_1,Q_2), Q_1, Q_2 are also not negative definite matrix. y_1 denotes the effect of demand change in the Grid Service Tier; and y_2 denotes the effect of demand change in the Admin Domain Tier. The more y_1 and y_2 are, the greater the effect is. The less y_1 and y_2 are, the smaller the effect is.

5 Stochastic Control Strategies

The offset system equation (7), parameters y_1 and y_2 in equation (9), (10) have described the effect that the change of users' demand arouses the change of resources requirement. We should select a control u_k in order to make the effect minimum. In particular, the users' demand is random, and the external disturbance is also random. Therefore, our work is how to select u_k, such that:

$$\min_{u_k} J = E\{\sum_{k=1}^{N-1}(x_k^T Q x_k + u_k^T u_k) + x_N^T Q x_N\} \tag{11}$$

where Q is not negative define matrix. The objective function means how to select the control, u_k, so as to keep the surplus resources and the demand offset to a minimum .

The formulas (7), (8), (11) denote state equation, measurement equation and performance index function respectively. This is a Linear Quadratic Gaussian (LQG) model [9]. According to separate principle, the problem may be divided into feedback

control and state estimate. The feedback state is Kalman Filter. In order to make equation (13) minimum, the optimal control is:

$$u_k = -F_k \hat{x}_{k|k-1} \tag{12}$$

where $\hat{x}_{k|k-1}$ denotes the estimate of state x_k. The feedback control gain is:

$$F_k = [I + B^T P_{k+1} B]^{-1} B^T P_{k+1} \tag{13}$$

$$\begin{cases} P_k = P_{k+1} + Q - P_{k+1} B_k [I + B^T P_{k+1} B]^{-1} B_k^T P_{k+1} \\ P_0 = Q \end{cases} \tag{14}$$

The state optimal estimate, $\hat{x}_{k|k-1}$, is:

$$\hat{x}_{k+1|k} = \hat{x}_{k|k-1} + B_k u_k + G_k [y_k - \hat{x}_{k|k-1}] \tag{15}$$

The gain of Kalman Filter is:

$$\begin{cases} G_k = S_{k|k-1} [R_2 + S_{k|k-1}] \\ S_{k+1|k} = S_{k|k-1} + R_1 - S_{k|k-1} [R_1 + S_{k|k-1}]^{-1} S_{k|k-1} \\ S_{1|0} = R_0 \end{cases} \tag{16}$$

where R_0 is the variance of random variable x_1.

Therefore, the surplus resources and the reserved resources of each tier of the hierarchical grid service system are:

$$\begin{cases} x_k^f = x_k + x_k^s \\ u_k^f = u_k + u_k^s \end{cases} \tag{17}$$

6 Simulation Analysis

In this section, we present results from a simulation study. The results show that by using the policy obtained from the stochastic control theory, we can not only maintain the system stability, but also reduce the variability caused by the input disturbance. As a result, the system can reduce the amount of uncertainty in the cost of hosting a grid service.

We assume the system has n=5 users, m=3 grid service nodes. In the noise condition, w_k submit to random normal distribution, i.e.

$w_k \sim N(0, \sigma_w^2)$, $\sigma_w^2 = 0.01$. The covariance matrix:
R_1=diag(0.0095,0.0088,0.0061,0.0071,0.0077,0.0102,0.0083,0.0124).
$v_k \sim N(0, \sigma_v^2)$, its covariance matrix :
R_2=diag(0.0083,0.0123,0.0072,0.0092,0.0063,0.0108,0.0080,0.0109).

The initial condition x_1^T =(0 1 1 0 0 0 1 0), its covariance matrix R_0=diag(0.05,0.05,0.05,0.05,0.05,0.1,0.1,0.1).
L=[0.2 0.4 0.2 0 0.1;0.6 0.5 0.8 1 0.8;0.2 0.1 0 0 0.1], k=7

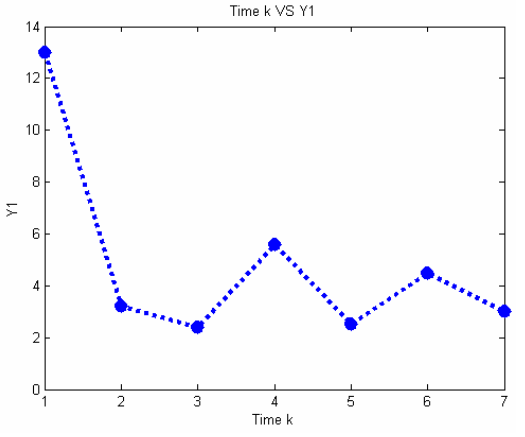

Fig. 2. The Curve of parameter y_1 with time k

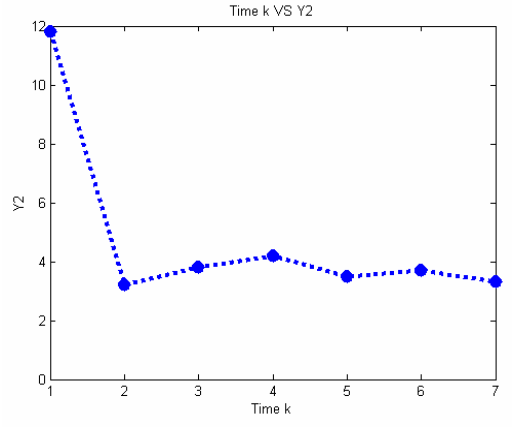

Fig. 3. The Curve of parameter y_2 with time k

Figure 2 and figure 3 show how the control strategy varies with the state. The parameter y_1 denotes the effect of the users demand disturbance in grid service tier, and y_2 does in AD tier. In both figures, x-axis presents time k, and y-axis presents y_1 or y_2. As shown in the plot, when time, k, increases, the effect factor of the system, y_1 or y_2, decreases. But the fluctuation in figure 2 is greater than that in figure 3. This shows that the effect subjected to in grid service tier is bigger than in the AD tier, i.e. the higher the tier is, the bigger the effect is. Therefore, we conclude that the stochastic control approach significantly reduces the disturbance of users demand of hosting of a dynamic grid service.

7 Conclusion

This work introduces a stochastic control model for deployment and hosting of a hierarchical grid service. The objective of the model is to produce policies to keep the costs to a minimum while maintaining the quality of service (QoS). The model is useful for making resource deployment decisions in the face of such uncertainties as random demand for the service. By employing a stochastic control approach, we obtain a solution of how to control the total cost in case of random users demand. When the users' demand produces a disturbance, the orders placed by the users will change. Meanwhile, this tier provides the downstream layer with users demand information. Furthermore, the disturbance will increase continuously from user tier to node tier. We quantify the effect by defining two parameters and produce a stochastic control algorithm. At last, a simulation experiment confirms that our mode can reduce the total cost and has a good performance.

Our work does not provide each tier of the hierarchical grid service framework with complete access to users demand information. When all layers of the framework share demand information, whether the effect still exists will be our future research topic.

References

1. I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. International Journal of Supercomputer Applications, 15(3), 2001.
2. I. Foster et al. Grid services for distributed system integration. Computer, 35(6), 2002.
3. D. England and J. B. Weissman, A Stochastic Control Model for the Deployment of Dynamic Grid Services, 5th. IEEE/ACM International Workshop on Grid Computing. 2004.
4. J. B. Weissman, S. H. Kim, and D. A. England. A Dynamic Grid Service Architecture. in submission, 2004.
5. J. B. Weissman and B. D. Lee. The service grid: Supporting scalable heterogenous services in wide-area networks. In IEEE Symposium on Applications and the Internet, 2001. San Diego, CA.
6. J. B. Weissman and B.-D. Lee. The virtual service grid: An architecture for delivering high-end network services. Concurrency: Practice and Experience, 14(4):287.319, Apr. 2002.
7. R. Buyya et al. Economic models for resource management and scheduling in grid computing. Concurrency and Computation: Practice and Experience, 14(13-15):1507.1542, 2002.
8. R. Wolski et al. Grid resource allocation and control using computational economies. In F. Berman, G. Fox, and A. Hey, editors, Grid Computing: Making the Global Infrastructure a Reality, chapter 32, pages 747~769. John Wiley and Sons, 2003.
9. Guo shanglai. A Stochastic Control, Tsinghua University, Beijing, 2000: 185~203.