

Feedback-Based Load Distribution for Web Clusters

Ying Lu[†], Asser N. Tantawi, Malgorzata Steinder, Giovanni Pacifici

[†] Dept. of Computer Science and Engineering
University of Nebraska - Lincoln
Lincoln, NE 68588
ylu@cse.unl.edu

Service Management Middleware Dept.
IBM T.J.Watson Research Center
Hawthorne, NY 10532
{tantawi, steinder, giovanni}@us.ibm.com

1 Introduction

Web clusters are the leading architectures for highly accessed web sites. They are usually composed of multiple server layers, for processing HTTP requests, implementing application functionalities and accessing databases. To manage the performance of such complex systems is challenging. It requires proper control mechanisms such as resource provisioning and load balancing. More importantly, those control mechanisms should collaborate with each other and work together harmoniously.

This poster presents a feedback-based load distribution mechanism for application layer servers in a web cluster. The contributions are twofold. First, the load distribution mechanism collaborates well with the dynamic resource provisioning. Second, it is designed based on feedback control theory to provide desired performance that is robust to workload changes.

2 Architecture

In a web cluster, requests belonging to the same application should be balanced among all servers that can process them so as to receive similar performance. The characteristics of numerous applications hosted by the web cluster may vary in a significant way in terms of processing requirements. Further an application instance may share the resources of a server node with variable other applications. A simple load balancing dispatcher, which attempts to equalize processor utilization among all nodes controlled by its gateway, does not yield equal performance to requests belonging to the same application, nor does it adapt to heterogeneous hosts properly.

We introduce a gateway component called WeightCalculator (Figure 1) which calculates load routing weights in a cell of heterogeneous server nodes, where each node is shared among multiple applications. The WeightCalculator periodically computes routing weights for each (application, server node) pair by properly reacting to the performance measures. The WeightCalculator is designed based on feedback control theory. It can guide the request dispatcher to balance the load despite the dramatic workload changes. The WeightCalculator also includes a mechanism that combines feedforward and feedback adaptations

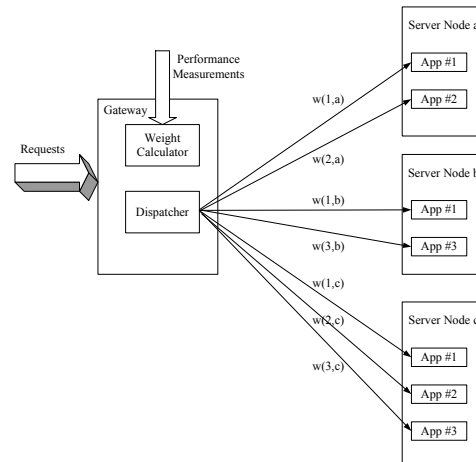


Figure 1. System Description.

to the dynamic resource provisioning, where the application placement or the number of server nodes is changed with workload fluctuations. In this poster, we will present how we adjust the structure of the feedback control system to deal with the application placement changes and how the WeightCalculator proactively and reactively adapts to the performance fluctuations caused by the placement changes. We demonstrate that by enabling the adaptations, the performance disturbances are significantly reduced.

The gateway component, WeightCalculator, has been implemented and evaluated in IBM WebSphere on demand operating environment. Web will describe the experience on putting control theory into practice. The problems and challenges of its deployment will be discussed.

3 Discussion

This poster presents our work on designing and implementing collaborative load balancing and dynamic resource provisioning in a web cluster environment. The proposed strategy is applicable to other similar scenarios, such as handling the interaction between the load distribution control and the server node power management. In general, for a complex system, interactions among its control components must be well-handled. How to apply feedback control theory to help design such complex adaptive systems will be an important research topic.