IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB7349

Chapter II

Network Service Availability and Performance

Mike Fisk
Los Alamos National Laboratory, USA

INTRODUCTION

As computer networks, specifically the Internet, become more and more integral to business and society, the performance and availability of services on the Internet become more critical. It is now a common need to provide a reliable network service to millions of Internet users and customers. The performance of these services is commonly a key factor in their success. Web portals and popular sites build relationships with customers based in part on their speed and availability. Even services internal to an enterprise frequently have serious consequences if there is a loss of availability.

This chapter discusses how advanced, multilayer switches can be used to increase the performance of network services. For this discussion, the term "performance" refers to availability, latency, and throughput, since all of these factors affect a user's impression of a site's performance. This chapter is intended for network service providers who must scale their services, network administrators who need to apply policies to their networks, and developers of switches who need to understand what the utility and requirements for these switches are. It is assumed that the reader has a working familiarity with networking principles, but substantial background information is also provided.

Clustering for Scalability and Availability

Scalability problems call for more capacity. This capacity can be had by upgrading existing systems, or by distributing load across multiple systems in a cluster. The clustering solution often incurs overhead for both management and

This chapter appears in the book, Enterprise Networking: Multilayer Switching and Applications by Vasilis Theoharakis and Dimitrios Serpanos.

operation, but is also more easily scaled to very large capacities. Purchasing large mainframes and supercomputers requires significant advance notice and capital. On the other hand, once a cluster has been created, it is often easy to incrementally add capacity to that cluster. Finally, clusters can use more mainstream, cost-competitive hardware than expensive systems that are suitable only for relatively small portions of the computing marketplace.

Availability problems can be caused by various problems, including hardware failure, natural disaster, software failure, attack, and ironically, success. The potential number of users of the Internet means that sites run the risk of becoming victims of their own success. Systems that may normally function fine may suddenly become unusable if there is dramatic change in their popularity.

Some threats to availability can be reduced through the use of fault-tolerant software and hardware, but there is always a level of problem severity that cannot be addressed. For many mainstream situations, extremely fault-tolerant hardware and software can be prohibitively expensive. As a result, clusters are also used to address availability concerns. As the number of systems in a cluster increases, the probability that they will all malfunction simultaneously grows exponentially. Of course, system architects must also take into account that the probability that every system will be functioning correctly also grows exponentially.

There are some kinds of availability and scalability problems that are very difficult to address without clusters. Most computer systems reside with a single machine room. If geographic distribution of resources are necessary, clustering is frequently the answer.

For all of these reasons, common trends in computing are leading towards increased usage of clusters and distributed systems. In this chapter, we discuss the techniques that can be used to let users access a clustered network service rather than a single machine.

To support this paradigm for network servers, it is critical that network application traffic can be directed to the best server. Very few network protocols have built-in support for the notion of multiple servers providing the same service. Rather than pushing this functionality into each application protocol, it is pragmatic to deploy the capabilities at a lower, infrastructural level.

To this end, many switches can now make switching and routing decisions based on application or service characteristics. Some of these characteristics, such as IP addresses, are used for traditional routing decisions, but others, such as HTTP URLs, TCP ports, and state information, have not been traditionally used.

Applying Policies to Networks

As computer networks become more utilized and more important, the management of this shared, limited resource becomes more important. It is frequently necessary to judge network traffic in terms of factors such as prioritization, fairness, and security.

For example, network traffic of unusually high or low priority may be segregated onto a different network connection. Or some traffic may be sent through