Available online at www.sciencedirect.com

ScienceDirect

Computer Networks 50 (2006) 3608–3621

Computer Networks

www.elsevier.com/locate/comnet

Optimal server allocations for streaming multimedia applications on the Internet

Padmavathi Mundur *, Poorva Arankalle

Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, United States

Received 3 May 2005; received in revised form 2 March 2006; accepted 6 March 2006
Available online 5 April 2006

Responsible Editor: D. Hutchison

Abstract

In this paper, we address the server selection problem for streaming applications on the Internet. The architecture we consider is similar to the content distribution networks consisting of geographically dispersed servers and user populations over an interconnected set of metropolitan areas. Server selection issues for Web-based applications in such an environment have been widely addressed; the selection is mostly based on proximity measured using packet delay. Such a greedy or heuristic approach to server selection will not address the capacity planning problem evident in multimedia applications. For such applications, admission control becomes an essential part of their design to maintain Quality of Service (QoS). Our objective in providing a solution to the server selection problem is threefold: first, to direct clients to the nearest server; second, to provide multiple sources to diffuse network load; third, to match server capacity to user demand so that optimal blocking performance can be expected. We accomplish all three objectives by using a special type of Linear Programming (LP) formulation called the Transportation Problem (TP). The objective function in the TP is to minimize the cost of serving a video request from user population x using server y as measured by network distance. The optimal allocation between servers and user populations from TP results in server clusters, the aggregated capacity of each cluster designed to meet the demands of its designated user population. Within a server cluster, we propose streaming protocols for request handling that will result in a balanced load. We implement threshold-based admission control in individual servers within a cluster to enforce the fair share of the server resource to its designated user population. The blocking performance is used as a trigger to find new optimal allocations when blocking rates become unacceptable due to change in user demands. We substantiate the analytical model with an extensive simulation for analyzing the performance of the proposed clustered architecture and the protocols. The simulation results show significant difference in overall blocking performance between optimal and sub-optimal allocations in as much as 15% at moderate to high workloads. We also show that the proposed cluster protocols result in lower packet loss and latencies by forcing path diversity from multiple sources for request delivery.

Keywords: Streaming multimedia; Video-on-Demand; Transportation Problem; Optimal server clusters; Quality of Service; Performance analysis

* Corresponding author. Tel.: +1 410 455 3019; fax: +1 410 455 3969.
E-mail addresses: pmundur@csee.umbc.edu (P. Mundur), apoorva1@csee.umbc.edu (P. Arankalle).

1389-1286/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2006.03.002

1. Introduction

Streaming applications for video such as pay per view and Video-on-Demand (VoD) have proliferated in the recent years spurred by the phenomenal increase, about 58% in the last year or two [1] in home-based Internet users. Industry-based streaming solutions consist of Real Networks' Real Media [2], Microsoft's Windows Media Services [3], Macromedia's Streaming Shockwave, and Apple's Quicktime Streaming [4]. Akamai [5] has moved closer to finding an efficient solution by using Content Delivery Networks (CDN) that put content closer to end users. However, we argue that moving the content to the edge alone will not work for multimedia applications where we use admission control as a mechanism to provide QoS. For such applications, capacity planning to realize better performance and resource utilization becomes a prominent issue. In this paper, we address system design issues to provide intelligent aggregation of server resources and request handling protocols that result in optimal blocking performance, server utilization and balanced loads. The analytical models and performance evaluation presented in this paper provide solutions to enhance the pragmatic approach provided by the CDNs. VoD-like applications require stricter QoS for longer duration than Web applications. Unlike Web servers, streaming servers are specially designed to deliver content in a predictable, delay-sensitive manner. These servers will use admission control and other related techniques to prevent overload and maintain QoS. Capacity planning to realize optimum server utilization in such an environment is the topic addressed in this paper.

In the spirit of CDNs, we consider metropolitan areas that are close together and content server nodes scattered throughout the area. We propose a special type of LP formulation called the Transportation Problem (TP) to obtain optimal server allocation to each metropolitan user population and aggregate those individual server nodes into clusters to serve each of the metropolitan areas. Organizing server clusters in this way provides distinct advantages: in an earlier work [31], we showed that a cluster of servers streaming different parts of the video to the client will result in better performance in terms of packet loss and latency. Using multiple sources that are physically separate, the data transfer is forced on multiple network paths and therefore, reduces the possibility of congestion. Employing a cluster of servers will also help reduce data granularity and achieve balanced load and fault tolerance. More importantly, organizing individual servers into clusters gives us a framework for monitoring blocking performance and detecting suboptimal server allocations when user demand changes.

We implement threshold-based admission control in individual servers within a cluster to enforce the fair share of the server resource to its designated user population. An arriving request will be rejected or *blocked* during admission control if there is no available capacity or if the number of requests from the user population exceeds its threshold. The blocking performance is used as a trigger to find new optimal allocations when blocking rates become unacceptable due to change in user demands and the current allocations are no longer optimal. In addition to providing such an architectural solution, we design and employ streaming protocols within server clusters for their efficient operation. The network context for our analysis is the best-effort IP-based networks. While we do not suggest any network related modifications, the proposed cluster architecture is expected to perform better by forcing path diversity in the network. Our work on streaming protocols and others suggest that such use of the network results in better performance in terms of packet loss and latency.

The TP formulation proposed in this paper is ideally suited for finding optimal server allocations for the server selection problem. The basic idea behind the TP formulation is to assign server capacities to user population demands by considering the cost of providing service from server i to a request from user population j. In mathematical programming problems where TP is applied, the cost is typically represented as distribution cost from manufacturing plant i to warehouse j. For our purposes, we use the round trip time (RTT) as the cost of serving a client request from server i to user population j. The network path metric RTT has long been used as a distance measure for proximity analysis in server selection for Web services. Using RTT as the cost metric in our TP formulation, the servers that are close to the user population are grouped together to provide service for that population.

The main contribution of this paper is the combined approach of providing a clustered server architecture based on optimal server allocation and protocols that efficiently operate those clusters. The proposed TP formulation in addition to addressing dynamic server selection problem, also