Dynamic Service Aggregation for Efficient Use of Resources in Interactive Video Delivery

D. Venkatesh and T.D.C. Little

Multimedia Communications Laboratory
Department of Electrical, Computer and Systems Engineering
Boston University, Boston, Massachusetts 02215, USA
{*dinesha,tdcl*}*@bu.edu*

Abstract. To support future interactive information delivery services there is a need to balance individual interactivity with the desire to maximize the number of supported sessions. Currently, few techniques have demonstrated the ability to renegotiate and scale service parameters per session in progress as required to adapt to differing terminal equipment characteristics and network congestion. This paper addresses this problem through the definition of decomposable service groups that permit aggregation of interactivity, terminal characteristics, and levels of service scaling. The proposed approach applies the characteristics of end applications and data storage requirements to the design of a data scaling mechanism.

1 Introduction

Evolving information delivery applications including video-on-demand (VOD), distance learning, and information browsing are becoming the dominant bandwidth consumers on the Internet and are expected to be so on future CATV and PSTN (public switched telephone networks). There is a trend among service providers towards supporting mixed bidirectional services. To support such services, broadcast technologies (e.g., CATV) must be modified support individual interaction by data recipients. Point-to-point technologies (e.g., Internet, PSTN) must be adapted to support multipoint data distribution. Fig. 1 illustrates such a spectrum of possible services for interactive video delivery. Case (a) requires a high-bandwidth outbound video stream from a data server and a low bandwidth interactive return signal. Case (b) is the multipoint scenario without interaction. Case (c) is fully interactive wherein each client requires a point-to-point session. This final case consumes the most system resources but is typical of true video-on-demand (T-VOD) services.

We propose a mechanism that effectively extends the number of viable sessions that a system can support. This mechanism is called *dynamic service aggregation*. In dynamic service aggregation, service scaling based on the level of interactivity is supported by providing a number of service groups, each permitting a different level of interactivity. Inter-group scaling involves shifting the user between different groups based on the level of interactivity demanded and the

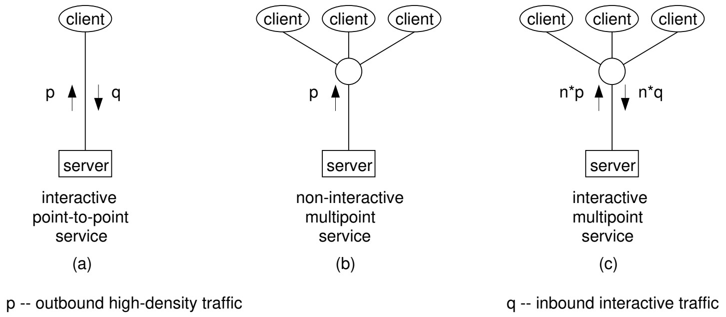

Fig. 1. Interactive Point-to-Point vs. Broadcast Services

availability of network resources. Sessions can be *promoted* by moving them to a group with a higher level of interactivity, or *aggregated* (demoted) by clustering them with similar sessions.

Consider the following distance learning scenario to illustrate the proposed mechanism: An instructor begins a lecture at 1:00PM. Due to limited server I/O bandwidth and network connection service, access to the course and course database is restricted to a window of opportunity during peak hours (e.g., mid-day). As the course is recorded on the fly, students can join a *service group* for the live multicast or can join late, forming a new group skewed by 10 min. Students who join late can "catch up" by starting interactive point-to-point sessions and browsing the recorded content and moving through it at a quicker pace. Ultimately they *aggregate* their interactive sessions by joining the live service group. When the instructor presents a reference to the previous class, some students may decide to access the video recorded from the previous class. This group can be either aggregated into a single session or spawned out as individual point-to-point connections, depending upon the availability of network resources. Near (in time) sessions are demoted by aggregation into service groups or a subset of the point-to-point sessions are provided with substandard service. A similar scenario exists for entertainment and information (news) video delivery, but with a much larger user population and potential for scaling gains.

The aforementioned scenario benefits from service scaling within and between service groups. The dynamic service aggregation is a greedy approach in that it attempts to reclaim network and server resources from interactive sessions by aggregating them with related non-interactive service groups when they become passive. The proposed scheme assumes clustering (in time) of user requests for common information items as is typical in large-scale video delivery scenarios. This assumption is most valid when it is most useful; as the clustering of requests will most likely occur during periods of high system utilization (e.g., when a new movie or newscast is released). The remainder of this paper describes the dynamic service aggregation protocol in more detail.

2 Dynamic Service Aggregation

Service Group Definition: *Service Groups* allow us to aggregate sessions with similar service (QOS) needs. A *Service Group* G_i is defined as a set of client-to-server sessions that can be characterized by a tuple

$$G_i = (T_j, I_l, S_k) \qquad T_j \in T, I_l \in I, S_k \in S,$$

where T_j represents a terminal-device class, I_l represents an interaction class, and S_k represents a QOS service class; each within their respective superclasses (i.e., T, I, S). I characterizes the delivery of a particular stream of information (e.g., the delivery of an instance of a video recording). The existence of identical tuples indicates the presence of nontrivial service groups (non-unitary) and the gains yielded via aggregation. Masking of terms permits aggregation of service groups across the masked attribute. This masking facilitates an appropriate group decomposition by the scaling mechanism for various services. The number of service groups at a given time is determined by the cardinality of the set of G_i (with masked terms dropped), or $|\cup G_i|$.

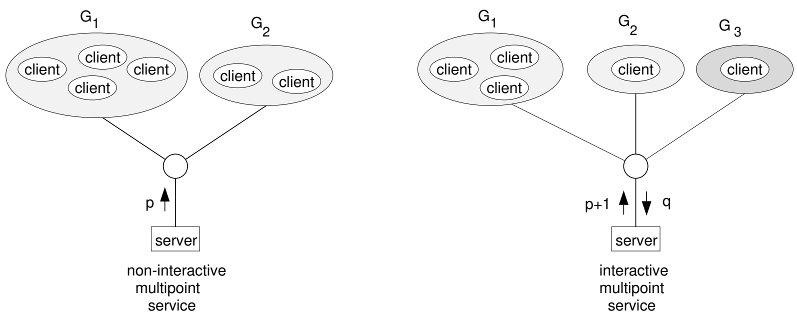

Fig. 2. Service Group Promotion

A service group with more than one element cannot be a T-VOD (interactive) session. A service group of size one represents the presence of a point-to-point interactive session. For broadcast services, users are initially assigned to a single service group based on their terminal and QOS requirements. When a user decides to shift from one mode of operation to another (e.g., broadcast to interactive), the system accommodates the switch by creating a new service group. The tuple describing the new service group differs by the I_l term. This example is illustrated in Fig 2.

Criteria for Service Promotion or Aggregation: Service promotion and aggregation is intended to provide interactive services to virtually all customers while simultaneously supporting a large customer population and yielding improved system utilization. However, this requires that the system carefully manage the distribution of resources. The scheme proposed in this paper attempts to reclaim unused interactive sessions (*passive* sessions) by aggregation. To do this we define the criteria for promoting or demoting sessions to service groups or the consolidation of service group as follows:

A passive session is defined as a session in which the user receives a data stream but has not interacted for some time period Δ. As a first criterion, we propose to aggregate passive connections that have a temporal locality within some duration ϵ. An appropriate value of ϵ is not known at this time; however, we expect this value to be in the range of seconds to minutes, depending on the constraints of the system resources and the number of supported clients. For example, existing PPV systems use values of order 30 minutes for ϵ. Clearly, skewed session groups cannot be merged arbitrarily for temporal locality within an ϵ as this can cause abrupt changes in the data stream and loss of continuity. However, the potential gains by using this approach are significant and are justified because they can be returned to the customer.

The criterion for promotion is the initiation of interaction by an existing client. The system can then try to accommodate this new interaction by the establishment of a dedicated point-to-point session. If appropriate, the newly defined service group can be immediately aggregated with an equivalent service group. In addition, if a client desires improved session quality, a session can be transferred to a different service group for a higher cost. This is achieved by manipulation of the S_k term. Finally, service demotion (aggregation) can be applied by the system to to reduce congestion or server loading.

3 Summary

We described a scheme to efficiently support multipoint information delivery applications by providing levels of constrained interaction and quality of service (QOS), and a means to switch among them. These levels are designed to allow flexibility for clients to adapt to interactive or non-interactive service at various levels of service quality, thereby permitting the system to accommodate the largest possible user population without sacrificing interactive functions.

References

1. Eleftheriadis, A., et al., "Multicast Group Address Management and Connection Control for Multi-Party Applications," Submitted for publication to the *IEEE/ACM Trans. on Networking*, 1994.
2. C. Szyperski and G. Ventre. "Efficient Group Communication with Guaranteed Quality of Service," *Proc. 4th IEEE Workshop on Future Trends in Distributed Computing Systems,* Lisboa, Portugal, September 1993.
3. Vonderweidt, G. et al., "A Multipoint Communication Service for Interactive Applications," *IEEE Trans. on Communications*, Vol. 39, No. 12, 1991, pp. 1875-1885.

This article was processed using the LaTeX macro package with LLNCS style