SomeCast

A Paradigm for Real-Time Adaptive Reliable Multicast*

Jaehee Yoon
jaeheey@cs.bu.edu

Azer Bestavros
bestavros@cs.bu.edu

Ibrahim Matta
matta@cs.bu.edu

Computer Science Department
Boston University
Boston, MA 02215

Abstract

SomeCast is a novel paradigm for the reliable multicast of real-time data to a large set of receivers over the Internet. SomeCast is *receiver-initiated* and thus scalable in the number of receivers, the diverse characteristics of paths between senders and receivers (e.g. maximum bandwidth and round-trip-time), and the dynamic conditions of such paths (e.g. congestion-induced delays and losses). SomeCast enables receivers to dynamically adjust the rate at which they receive multicast information to enable the satisfaction of real-time QoS constraints (e.g. rate, deadlines, or jitter). This is done by enabling a receiver to join Some number of concurrent multiCast sessions, whereby each session delivers a portion of an encoding of the real-time data. By adjusting the number of such sessions dynamically, client-specific QoS constraints can be met independently. The SomeCast paradigm can be thought of as a generalization of the AnyCast (e.g. Dynamic Server Selection) and ManyCast (e.g. Digital Fountain) paradigms, which have been proposed in the literature to address issues of scalability of UniCast and MultiCast environments, respectively.

In this paper we overview the SomeCast paradigm, describe an instance of a SomeCast protocol, and present simulation results that quantify the significant advantages gained from adopting such a protocol for the reliable multicast of data to a diverse set of receivers subject to real-time QoS constraints.

Keywords: Reliable Multicast, Real-Time Communication, FEC using Reed-Solomon-like codes, Simulation.

*This work was supported in part by NSF research grants ESS CCR-9706685, CAREER ANIR-9701988, and MRI EIA-9871022.

1 Introduction

The ubiquity and acceptance of the Web in our society has encouraged the development of many Internet applications that are inherently of a real-time nature—or that deal with information that is inherently temporal in nature. The communication of real-time information over the Internet is challenging due to the inherent unpredictability involved in using such an open infrastructure. This unpredictability is documented in a number of studies that confirm the highly variable nature of Internet traffic over a multitude of time scales [10, 12, 13], and the futility of techniques such as buffering to eliminate such variability [32].

Motivation: An important class of real-time applications requires the communication of the same content to a very large number of receivers. In order to cope with the highly-variable nature of network congestion, applications often trade reliability for timeliness, or trade timeliness for reliability.

Trading reliability for timeliness is a common practice that involves the use of a rate-based transport (e.g. using UDP for the communication of audio/video streams). An example of this approach is the MBone multicast protocol. This approach results in the deployment of congestion-insensitive applications, and is viewed by the Internet community as a bad practice [26]. Even if acceptable, such an approach would only be useful for real-time Internet applications that are able to tolerate a degree of unreliability (i.e. packet losses). For many real-time applications, such unreliability is not tolerable. Examples include group simulations, live auctions, real-time content replication for Web portals, and stock brokerage applications. Such applications require a multicast infrastructure that is both *real-time* and *reliable* [40].

A common approach to improving reliability is through the use of redundancy. Two forms of redundancy are typically exploited: temporal and spatial. Temporal redundancy involves the use of the same set of resources to recover from failures over an extended period of time. An example of this approach is the use of ARQ-based (or retransmission-based) recovery techniques. Spatial redundancy involves the use of additional resources to mask failures. An example of this approach is the use of FEC techniques.

Trading timeliness for reliability is evident in all communication protocols that use retransmissions to recover from packet losses due to network congestion (including TCP). An example of this approach for multicast communication is the Scalable Reliable Multicast (SRM) [15] and the Cyclic UDP Multicast [2] and the Digital Fountain Multicast [7] paradigms. Obviously, such techniques are not adequate for real-time applications, for which “a late packet is a lost packet”.

Trading resources for reliability is common for mission-critical systems that cannot tolerate recovery delays (e.g. using N-modular redundancy for collision avoidance systems) or for system components with irrecoverable failure modes (e.g. using mirror disks to protect against a disk crash).

While a widely accepted practice for hard real-time systems, the trading of resources for reliability and/or timeliness is not a common practice for systems with “softer” deadline constraints, for example over the Internet. In this paper, we argue that the use of redundant resources to improve the reliability and timeliness of Internet applications in general (and multicast communication in particular) is appropriate due to the *already existing* multiplicity of resources in such systems—a multiplicity that is required for performance and scalability purposes.

Contribution: In this paper we present SomeCast—a multicast paradigm that enables the satisfaction of timing constraints without sacrificing communication reliability. This is done by enabling receivers to dynamically adjust the rate at which they receive multicast information to guarantee the satisfaction of real-time QoS constraints (e.g. rate, deadlines, or jitter). This rate adjustment is made possible by enabling a receiver to join Some number of concurrent multiCast sessions, whereby each session delivers a portion of an encoding of the real-time data. By adjusting the number of such sessions dynamically, client-specific QoS constraints can be met independently. The SomeCast paradigm can be thought of as a generalization of the AnyCast and ManyCast paradigms, which have been proposed in the literature to address issues of scalability of UniCast and MultiCast environments, respectively. The SomeCast is *receiver-initiated* and thus scalable in the number of receivers, the diverse characteristics of paths between senders and receivers (e.g. maximum bandwidth and round-trip-time), and the dynamic conditions of such paths (e.g. congestion-induced delays and losses).

Scope: We start this paper with a review of related work in Section 2. We follow that with an overview of the SomeCast paradigm in Section 3. We present an instance of a SomeCast-based real-time reliable multicast protocol in Section 4. We present performance evaluation results that quantify the benefits of our proposed protocol in Section 5. Finally, we conclude with a summary and an overview of future work in Section 6.

2 Related Work

ARQ-based Techniques: The first category of reliable multicast transport protocols is based on ARQ (Automatic Repeat reQuest). Here, the sender retransmits lost data upon request from the receiver. A straightforward application of ARQ in a multicast setting results in the so-called *NACK implosion* problem. This problem occurs when every receiver sends a NACK message to request retransmission of the same packet, causing an implosion at the sender. To prevent this implosion of control packets, Xpress Transport Protocol (XTP) [41] proposed that a receiver multicasts control packets to the entire group. A receiver waits for a random time before sending a NACK packet, and refrains from sending a NACK if it sees a NACK from another receiver for the same packet. SRM (Scalable Reliable Multicast) [15] uses similar mechanisms to control the sending of request (NACK) and repair (retransmitted data) packets. In SRM, the random delay before sending a request (repair) packet is a function of the receiver's distance from the node that triggered the repair (request). Each node estimates its distances from other nodes by multicasting session messages. The random timer is set to be inversely proportional to the distance. Thus, although a number of receivers may all miss the same packet, a receiver close to the point of failure is likely to timeout first and multicast the request. Other receivers that are also missing the data hear the request and suppress their own request.

As we hinted earlier, retransmission-based approaches trade timeliness for reliability and hence are not useful for the class of multicast applications that require *both* reliability and timeliness.

Deadline-Cognizant Techniques: Most reliable transport protocols (whether unicast or multicast) are incognizant of the temporal semantics of the data being communicated. As a result, a reliable transport protocol may end up wasting resources attempting to recover from the loss of a packet that is of no value to the receiver (because it is late). Such wasteful utilization of resources may result in delaying more packets, resulting in further violations of timeliness constraints. The work of Li, Ha, Varghavan [21] is an example of an approach that attempts to address such a scenario. While this technique was proposed primarily for unicast communication (namely TCP), it is conceivable that similar techniques could be used to avoid unecessary retransmissions in a multicast environment (e.g. SRM).

Deadline cognizance is likely to improve the timeliness of a reliable communication by preserving network resources, but it is unable to mask (or hide) the delays resulting from congestion between a sender and a receiver.

FEC-based Techniques: Another category of reliable multicast protocols is based on FEC (Forward Error Correction). Here, the original data is encoded to obtain additional repair packets that are used to recover data packet loss. An example of this approach is the use of FEC in the SHARQFEC protocol of Kermode [20] and the use of FEC in the real-time reliable multicast of Rubenstein, Kurose, and Towsley [38].

Compared to retransmission-based approaches, FEC-based techniques enable a more efficient (and timely) recovery from packet losses. However, they cannot mask (or hide) the delays resulting from congestion between a sender and a receiver.

Multi-Layer-based Techniques: One approach to providing scalable reliable multicast is the use of multiple channels (or layers), whereby receivers experiencing a higher degree of losses join more channels to recover lost packets. An example of this approach is the work of Kasera, Hjalmtysson, Towsley, and Kurose [19].

Multi-layer-based techniques—while effective in dealing with the variability of loss characteristics across a large set of receivers—do not allow receivers experiencing high loss rates to recover from such losses in a timely fashion. In other words, the reliability of a multicast is guaranteed, but not its timeliness.

AnyCast-based Techniques: To adapt to the dynamic conditions of the network, several techniques have been proposed to enable a receiver to select "the best" one of many possible servers to fulfill its request. Such selection could be done at the server or at the client. An example of a server-based approach is the *AnyCast* paradigm of Fei, Bhattacharjee, Zegura, and Ammar [14]. An example of a client-based approach is the dynamic server selection protocol of Carter and Crovella [9]. While these techniques were proposed primarily for unicast communication, it is conceivable that a similar AnyCast paradigm could be used to select the "best" multicast group for a given receiver (out of many possible alternatives offering the same service).

Given the high variability in network conditions, AnyCast-based techniques are unlikely to be effective for applications that involve a prolonged communication session (e.g. video-on-demand).

ManyCast-based Techniques: Another approach to speeding up access to popular content is the use of multiple sources concurrently—or *ManyCasting.* This is exemplified in the work of Byers, Luby, and Mitzenmacher [8], which use Tornado encoding to ensure the efficiency of the retrieval and reconstruction processes. While not addressing the problem of real-time reliable multicast specifically, this work is similar to ours in that it enables a receiver to communicate concurrently with many senders.

3 Overview of the SomeCast Paradigm

From AnyCast and ManyCast to SomeCast: The SomeCast paradigm can be thought of as a generalization that encompasses both the AnyCast and the ManyCast paradigms to which we eluded in Section 2.

Both the AnyCast and the ManyCast paradigms are *server-based* approaches for improving the scalability, fault tolerance, and performance of *best-effort* Internet systems (e.g. Web servers). Under the AnyCast paradigm, the "best possible" provider of service is identified out of many such providers of service. Under the ManyCast paradigm, "all" providers of service are contacted to speed up the service. In both cases, the multiplicity of providers of service is necessary to deal with the issue of scale in a *best-effort* environment.

Under the SomeCast paradigm, clients are empowered to exploit the multiplicity of resources so as to meet specific reliability and real-time QoS constraints. In other words, the SomeCast paradigm is *client-based.* It enables multiple resources available in an inherently best-effort environment to be leveraged to achieve a prescribed QoS. A client (or receiver) contacts "some" providers of service as needed. The delegation of QoS management to clients is attractive because it enables the receivers to have very diverse QoS requirements without resulting in a state-explosion problem at the sender (or network).

Architecture of Content Delivery under SomeCast: Figure 1 illustrates the general architecture of a SomeCast system. We assume that "content" is to be delivered from a *source* to a potentially very large number of *receivers* (or clients) through a number of *senders*, each of which acts as a proxy of the source (i.e. as an outlet for the content).[1] Under SomeCast, each sender sets up a multicast group and a receiver joins as many multicast groups as *necessary* to satisfy its QoS constraints, in an adaptive fashion.

It is important to emphasize that there are *two* "distribution" problems in the architecture

[1]This architecture—comprising a large number of servers acting as proxies for a single content source—is quickly being accepted as inevitable for scalable Internet systems [1, 39, 18].

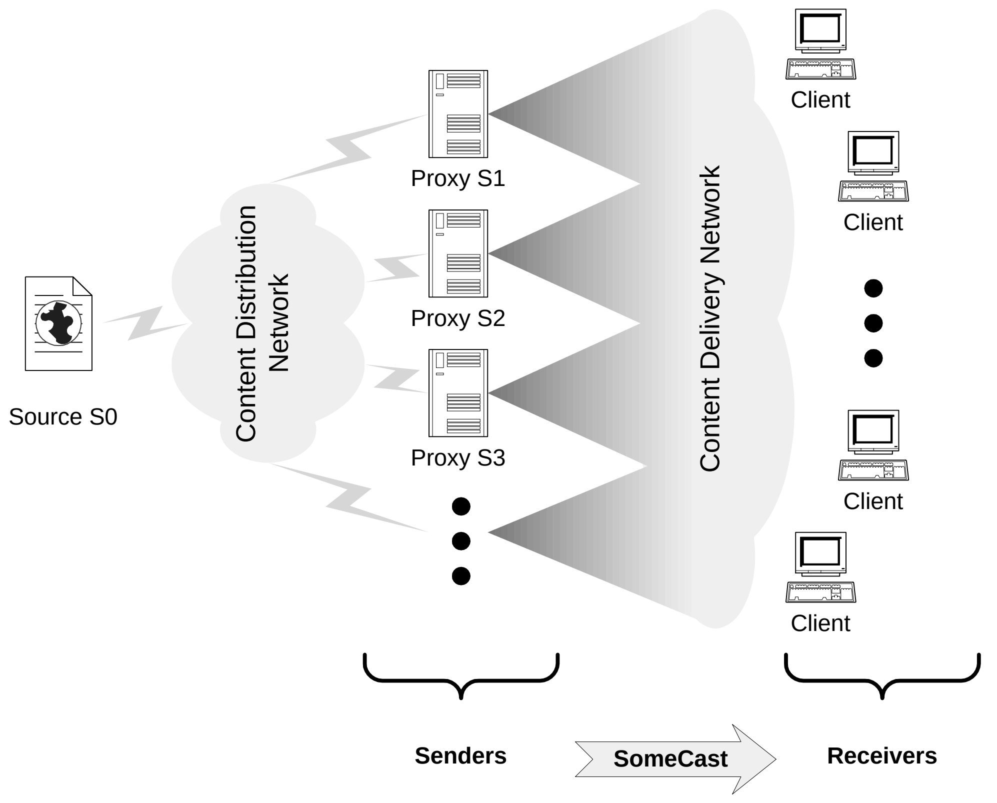

Figure 1: The SomeCast Paradigm

depicted in Figure 1: (1) the source must distribute the content to its proxies (the senders), and (2) the senders must relay that content to the receivers. These two problems are quite different. In the first, the system is "closed" in the sense that the set of proxies are all known *a priori*, and are (most likely) within the confines of a single organization or network (e.g. content replication services on the Internet [1, 39]). In the second, the system is "open" in the sense that the (potentially very large) set of receivers are not known *a priori*; they operate independently and may require significantly different QoS. The SomeCast paradigm addresses the challenges posed by the second of the above two distribution problems.[2]

Store-and-Forward versus Streaming Multicast Models: Another important consideration in the architecture depicted in Figure 1, is the nature of the content being distributed. Two possibilities exist: (1) the content comprises a single object (e.g. current bids) that is updated frequently by the source, or (2) the content comprises a live feed (e.g. prices on the stock exchange) that is constantly generated at the source. The SomeCast paradigm can be used to support *both* of these models.

Under the first model, senders act as *repeaters*. They continuously and repeatedly multicast the most-up-to-date content on their respective multicast groups. Receivers join as many such multicast groups as necessary to retrieve such content in a timely fashion. The continuous, periodic

[2]There are many products in the market-place that address the distribution of content from source to proxies [18, 40].

retransmission of content by senders is similar to the Broadcast Disk techniques proposed in [3, 4], the Digital Fountain techniques proposed in [7], and the Cyclic Best Effort UDP Protocol proposed in [2]. Thus, under this model, content flows in a store-and-forward fashion from the source to the senders and then from the senders to the receivers. Example applications that fit that model would be the multicast of radar information, or the multicast of the status of an on-line auction.

Under the second model, senders act as *relays*. They receive a segment of the stream, which they multicast once on their respective multicast groups. Receivers join as many such multicast groups as necessary to be able to recover such segments in a timely fashion (i.e. before the senders switch to the next segment). Thus, under this model, content flows in a pipelined fashion from the source to the senders to the receivers. Example applications that fit that model would be the multicast of live feeds from a stock market exchange, or the multicast of sensory information or live video.

It should be clear that from the perspective of distributing content from the proxies (i.e. senders) to the receivers, the problem is identical under both models. Thus, to simplify our presentation, for the rest of this paper (and without loss of generality), we assume that the first of the above two models is in play.

Reed-Solomon Encoding in SomeCast: In SomeCast, receivers receive the multicast content from multiple senders. Thus, a key component of the SomeCast paradigm is the use of a mechanism that ensures that the various segments of content (received from the various senders) are independent, and thus can be combined efficiently to obtain the original content. To that end, SomeCast assumes that content is encoded using a Reed-Solomon encoding mechanism.

Reed-Solomon Codes (RSC) [27] are a popular FEC coding technique, which is used in many FEC-based reliable multicast protocols [20, 38]. Reed-Solomon codes are based on the arithmetic of finite (Galois) fields [35]. Reed-Solomon-like codes have been proposed and used in a number of projects for efficient information retrieval. Examples include (1) the Information Dispersal Algorithm (IDA) [34] used for efficient, secure, and fault-tolerant parallel data access [25], (2) the Adaptive IDA communication protocol [5] used in TCP Boston to address the fragmentation problem of IP over ATM [6], and (3) the Tornado codes [23] used in Digital-Fountain multicast [7].

Conceptually, a Reed-Solomon code is a mapping from an m-dimentional vector space over a finite field K into a vector space of higher dimension over the same field. Starting from a data segment $(s_0, s_1, ..., s_{m-1})$, where each s_k is an element of the field, a Reed-Solomon code produces $(P(0), P(g), P(g^2), ..., P(g^{N-1}))$, where N is the number of elements in K, g is a generator of the cyclic group of nonzero elements in K, and $P(x)$ is the polynomial $s_0 + s_1 x + ... + s_{m-1} x^{m-1}$. If N is greater than m, then the values of P overspecify P, and the properties of finite fields guarantee that the coefficients of P (namely, the original data segment) can be recovered from any m of the values.

The SomeCast paradigm we propose in this paper is *independent* of the specific Reed-Solomon coding technique chosen for an implementation. However, to make our presentation concrete—and for purposes of illustration and derivation of specific realizations—we will adopt one such coding technique, namely the Information Dispersal Algorithm (IDA) of Rabin [34]. Note that other Reed-

Solomon-like coding techniques (e.g. Tornado [23]) may have more attractive properties than those of IDA with respect to encoding and decoding efficiency (for example). Thus, if SomeCast is to be deployed, IDA may not be the best choice. Nevertheless, for the purposes of this paper, IDA's elegance and simplicity will allow us to focus on the details of the SomeCast paradigm as opposed to the details of the underlying encoding/decoding technology.

To understand how IDA works, consider (a segment of) a data object to be multicast. Let that object consist of K *blocks* (or packets). Using IDA's *dispersal operation*, this object could be processed to obtain $s * K$ blocks, where $s > 1$ is the *stretch factor*. Recombining *any* K of these blocks, using IDA's *reconstruction operation*, is sufficient to retrieve the original data object. Figure 2 illustrates the dispersal, communication, and reconstruction of an object using IDA.

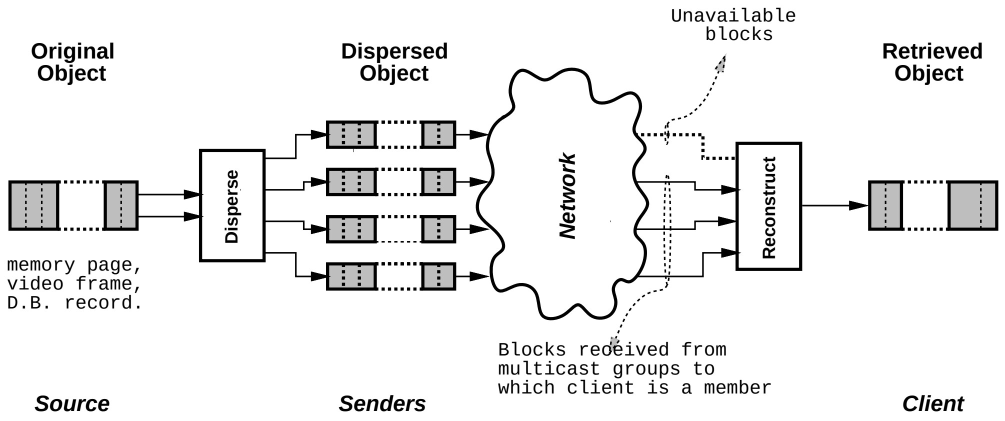

Figure 2: Dispersal and reconstruction of information using IDA.

4 Real-Time Reliable Multicast using the SomeCast Paradigm

In this section, we present an instance of a SomeCast-enabled protocol that empowers a set of receivers to satisfy diverse real-time QoS constraints in a reliable multicast setting. In Section 4.1, we give an overview of the protocol, followed in Section 4.2 with a detailed description.

4.1 Protocol Overview

We consider the reliable multicast of an object of size K packets. We assume that the object is encoded so that each sender S_i has $u * K$ packets, where $u = s/\mathcal{S}$, s is the stretch factor of the encoding and $\mathcal{S}$ is the total number of senders. S_i starts a multicast group over which its packets will be sent only if there is at least one receiver that is a member of its group. At any point in time, if S_i finds that its group is empty (i.e., with no members), S_i stops transmitting packets.

Let S_0 be the primary sender. If there is no deadline requirements or the delay bounds are very loose, then a receiver may only join the multicast group of S_0 to receive at least K packets and reliably recover the original data by the deadline.

Initially, a receiver may join one or more multicast groups associated with one or more senders.

Those senders send packets over their multicast groups. By allowing a receiver to join all groups, the protocol can effectively handle stringent delay bounds. For less stringent delay bounds, the receiver may leave all but the first group associated with the primary sender S_0.

Periodically, a receiver updates its estimates of its loss rate and throughput of the path from sender S_i. A receiver can then estimate the total number of packets that it *expects* to receive by the deadline from senders/groups to which it is subscribed. If this is not enough to receive all K packets by the deadline, then the receiver joins as many groups as needed. On the other hand, if the receiver anticipates to receive *much* more than K packets by the deadline, then the receiver unsubscribes from "some" groups to only receive what is needed and thus avoid unnecessary transmissions. Without loss of generality, we henceforth assume that if a receiver decides to unsubscribe from some multicast groups, it does so by unsubscribing from higher numbered ones first.

In addition to dynamically adjusting the number of groups to which a receiver subscribes, in response to probes from senders, a receiver sends sender S_i NACK messages specifying how many more packets it needs from S_i. Thus, S_i can send the right amount of additional packets so that all receivers subscribed to its group meet their deadlines. Whenever a receiver receives K packets, it leaves all multicast groups to which it is currently subscribed.

4.2 Protocol Description

Table 1 introduces the notation we adopt throughout this paper to describe our SomeCast-enabled protocol. We describe the details of our SomeCast-enabled protocol by presenting the steps undertaken by the Sender(s) and Receiver(s) at various stages of the protocol.

Symbol	**Meaning**
K	The number of original data packets. Each receiver should receive K packets by the deadline
$\mathcal{S}$	Total number of senders
S_i	Sender S_i
s	Stretch factor of the Reed-Solomon-like encoding (e.g. IDA). Thus, the K original data packets are encoded to obtain up to $s * K$ packets.
$u * K$	The maximum number of packets transmitted by a sender, where $u = s/\mathcal{S}$. In this paper we take $u = 2$.
$maxseqno_i$	The maximum sequence number of data to be transmitted by S_i
$seqno_i$	The sequence number of packet/probe sent by/received from S_i
PTS_i	The number of packets yet-to-be-transmitted at the time S_i sends a probe
RPC_i	The Received Packet Counter denotes the number of packets received thus far from S_i
$TRPC$	The Total Received Packet Counter ($TRPC$) denotes the total number of packets received thus far
g	Current number of groups to which a receiver is subscribed
D	A receiver's deadline
R_i	Estimated throughput from sender S_i
L_i	Estimated loss rate on the path from sender S_i
RTT	The maximum Round-Trip Time between the sender and a receiver

Table 1: Notation used in our SomeCast-enabled protocol description

Sender: Start

SS.1 Each sender S_i $(i = 0, 1, \cdots, \mathcal{S} - 1)$ sets its initial $seqno_i$ of its first packet and its $maxseqno_i$ as follows:

$$\begin{aligned} seqno_i &= 2iK \\ maxseqno_i &= 2iK + K - 1 \end{aligned}$$

SS.2 If sender has receivers in its multicast group, it starts to transfer the first K data packets.

SS.3 Concurrently with step SS.2,[3] the sender applies coding to the first K data packets to obtain $2K$ packets.

Sender: Probing

SP.1 Periodically, a sender transmits a probe piggy-backed on a data packet. The probe consists of a time-stamp that identifies the time at which the probe is sent and PTS_i. Namely,

$$PTS_i = maxseqno_i - seqno_i$$

where $seqno_i$ is the sequence number of the packet transmitted with the probe.

Sender: NACK Processing

SN.1 Upon receipt of a NACK, sender S_i updates $maxseqno_i$ to the *maximum* requested by all receivers in response to the same probe, so as to make sure all receivers subscribed to the multicast group of S_i receive the additional packets they need by their deadlines.

SN.2 Upon receipt of a NACK, sender S_i also updates its estimate of the maximum Round Trip Time (RTT).

SN.3 If sender S_i does not receive NACK requests for 2 probing intervals to either increase or decrease $maxseqno_i$, then the bottleneck receiver which requested the current $maxseqno_i$ may have left the multicast group of S_i. S_i then decreases $maxseqno_i$ as follows:

$$maxseqno_i = seqno_i + (maxseqno_i - seqno_i)/2$$

SN.4 While transmitting the last RTT worth of packets, sender ignores NACK requests for decreasing $maxseqno_i$ to avoid errors due to possible under-estimation of losses in the last stage.

Sender: End

SE.1 Sender S_i stops its transmission once all members (receivers) in its group leaves.

[3]Our protocol may overlap data transmission with encoding/decoding, hence dramatically reducing latency for all receivers to receive the number of packets needed for recovering the original data by the deadline.

Receiver: Packet Processing

RP.1 Whenever a receiver receives a packet from sender S_i, it increases the Received Packet Counter (RPC_i) by one to keep track of the number of packets received from S_i. It also increments by one the Total Received Packet Counter ($TRPC$).

RP.2 If $TRPC$ is greater than or equal to K, the original data can be reconstructed from the packets received so far. The receiver then decodes the received data and leaves all multicast groups it is a member of.[4]Also, if the deadline has expired, receiver leaves all multicast groups it is a member of after dropping all (useless) packets it has received so far.

RP.3 If $TRPC$ is less than K and if the packet received is a probe, then the receiver proceeds as follows:

RP.3.1 Estimate the total number of packets, N, that it expects to receive by the deadline as follows:

$$N = \sum_{i=0}^{\mathcal{S}-1} RPC_i + \sum_{i=0}^{g-1} R_i \times (D - t)$$

where $\mathcal{S}$ is the total number of senders, g is the current number of groups to which the receiver is subscribed, D is the receiver's deadline, t is the current time, and R_i is the estimated throughput from sender S_i (computed periodically by the receiver in RE.3 below). The firm term in the right-hand side represents the number of packets already received, and the second term represents the number of packets that the receiver *anticipates* to receive by the deadline.

RP.3.2 Compute $maxseqno_i$ for each sender S_i the receiver is listening to (*i.e.*, receiver is currently a member of S_i's multicast group) as follows:

$$maxseqno_i = \min(seqno_i + \frac{R_i}{1 - L_i} \times (D - t),\ 2\,(i+1)\,K - 1)$$

where $seqno_i$ is the sequence number of S_i's probe, L_i is the estimated loss rate on the path from sender S_i, and t is the current time. Thus, the term $\frac{R_i}{1-L_i} \times (D - t)$ represents the number of *additional* packets S_i needs to send so that the receiver receives the number of packets it expects from S_i by the deadline. Note that $maxseqno_i$ can not exceed $(2\,(i+1)\,K - 1)$ since each sender is assumed to hold $2K$ encoded packets.

RP.3.3 If ($seqno_i + PTS_i < maxseqno_i$), then the forthcoming packets from S_i are not enough to recover the original data. The receiver sends a NACK that includes the new lower bound on $maxseqno_i$ calculated in step RP.3.2.

RP.3.4 If ($seqno_i + PTS_i \geq maxseqno_i$) and ($seqno_i + PTS_i$) equals a $maxseqno_i$ the receiver had sent to S_i in a previous NACK, then the receiver was the bottleneck and is now uneccessarily requiring S_i to send more packets than needed. The receiver sends a NACK that includes the new lower $maxseqno_i$ calculated in step RP.3.2.

Receiver: Periodic Estimation

RE.1 Periodically, a receiver updates its estimates of its loss rate l_i on the path from sender S_i to which it is subscribed as follows:[5]

$$l_i = 1 - (\Delta RPC_i / \Delta seqno_i)$$

where $\Delta seqno_i$ is the difference in sequence numbers of packets received from sender S_i at the beginning and end of the update time interval. ΔRPC_i is the number of packets received from S_i during the update interval. Thus, the ratio l_i gives the current proportion of S_i packets lost.

RE.2 Periodically, a receiver also updates its throughput r_i from sender S_i as follows:

$$r_i = \Delta RPC_i / \Delta t$$

where Δt is the length of the update interval.

RE.3 Based on l_i and r_i, a receiver maintains exponential moving averages and deviations of the loss rate and throughput for each sender S_i. Specifically,

$$\begin{aligned} AvgR_i &= (1-\alpha)\ AvgR_i + \alpha\, r_i \\ DevR_i &= (1-\delta)\ \mid r_i - AvgR_i \mid + \delta\, DevR_i \\ R_i &= AvgR_i + \gamma\, DevR_i \end{aligned}$$

where R_i is the estimated throughput from sender S_i. $AvgR_i$ and $DevR_i$ are the moving average and deviation, respectively.[6]

Similarly, the loss rate L_i from sender S_i is estimated.

RE.4 The receiver decides whether it should join or leave multicast groups based on N computed as in RP.3.1, i.e. the total number of packets that the receiver expects to receive by its deadline from all groups to which it is currently subscribed. If N is less than K, then the receiver joins as many groups as needed for N to exceed $K + relax$. $relax$ is a protocol parameter chosen to ensure that N is well beyond K.[7]On the other hand, if N exceeds $K + relax$, then the receiver unsubscribes from higher numbered groups for N to be just beyond $K + relax$ and thus avoid unnecessary transmissions.

Figure 3 shows the pseudo-code for the sender and receiver agents.

[4]By allowing receivers to leave multicast groups once they receive the K packets needed, we significantly reduce the bandwidth consumed over the network.

[5]In our experiments, we take the update period to be 0.2 seconds.

[6]In our experiments, we take $\alpha = \delta = 0.5$, and we set γ to 1. γ could be set to higher values to account for high variability in the case of stringent deadlines.

[7]In our experiments, we take *relax* to be 20 packets for $K = 1000$ packets.

4.3 Sender Probing and RTT Estimation

During data transmission, sender S_i transmits probes periodically. A probe packet includes the number of packets to be sent (PTS_i) and time-stamp for when the packet is sent. The purpose of using probes is two-fold: First, a probe is used to trigger NACKs from receivers. Upon receiving the probe containing PTS_i, a receiver makes a local decision whether this is enough to sustain its current loss rate as we described earlier in RP.3. Second, a probe is used to estimate the round-trip time (RTT). When a receiver responds with a NACK, it sends the time-stamp for when the NACK is sent. The time-stamps are used to calculate RTT in the same manner as in [28]: the sender sends a probe at time t_1, and a receiver receives the probe at time t_2. If the receiver sends a NACK at time t_3, it includes (t_1, Δ), where $\Delta = t_3 - t_2$. Once the sender receives the NACK at time t_4, it computes RTT as $RTT = t_4 - t_1 - \Delta$.

In other reliable multicast protocols such as SRM [15] and SHARQFEC [20], the estimation of RTT is very critical for NACK suppression and repair. However, in our protocol, the RTT value is not critical. The sender only needs an estimate of the maximum RTT from receivers to set the period of probing. As more data is transmitted, the feedback from receivers about their losses in response to probes becomes more critical. Therefore, it is desirable to set the period of probing to be large at first, and then gradually decrease it. However, the period of probing should be at least equal to RTT to avoid sending duplicate probes.[8]

5 Performance Evaluation

In this section we present the results of our prototype implementation and performance evaluation of our SomeCast-enabled protocol.

5.1 Simulation Model

To evaluate the performance of SomeCast, Unicast and ManyCast, we set up a simulated multicast network using the 19-node tree topology depicted in Figure 4. In this topology, a CBR (Constant Bit Rate) data source is attached to each of nodes 14 to 17 (the primary sender S_0 is attached to node 14). All other nodes (i.e. nodes 0 to 13) act as receivers. In our simulations, the packet interarrival time for the CBR source is set to 0.01 seconds. Each link in the network is subjected to a maximum of 32 on-off cross connections generated by a UDP-based agent. This UDP-based agent generates connections with an inter-arrival time uniformly distributed between 0 and 0.1 second. Each connection is an on-off source with Pareto distributed “on” and “off” periods with average durations of 0.1 second and 0.9 second, respectively. The Pareto distribution has a skew parameter of 1.35. During the “on” periods, packets are generated at a rate of 1000Kbps. This cross-traffic resulted in up to 30% loss rates observed at receivers. The bandwidth of the links in our simulated topology are set to 1.5Mbps. All links have a propagation delay of 15ms. The packet size is 1KB. We take $K = 1000$ packets, so the size of the data is 1MB.

[8]In our simulations, sender S_i sends the first probe after sending the first $K/5$ packets, then the probing frequency is increased by sending one probe every RTT.

```
initialize $maxseqno_i \leftarrow K-1$;
while $seqno_i \leq maxseqno_i$
    transmit packet;
    encode data into $2K$ packets;

    if $seqno_i$ = next probe then
        if no NACK arrived in last $2RTT$ then
            $maxseqno_i \leftarrow seqno_i +$
                $(maxseqno_i - seqno_i)/2$;
        $PTS_i \leftarrow maxseqno_i - seqno_i$;
        send probe with data;
    else
        send data;

    if NACK with new $maxseqno_i$ arrives then
        update $maxseqno_i$;
```

(a) Sender algorithm

```
initialize $TRPC \leftarrow 0$;
    if data packet is received then
        $TRPC \leftarrow TRPC + 1$;

        if $TRPC < K$ and $t \leq D$ then
            if probe arrived then
                compute loss rate $L_i$ and throughput $R_i$;
                estimate $maxseqno_i$ needed from sender $S_i$;
                if $PTS_i + seqno_i < maxseqno_i$ then
                // receiver needs more additional packets
                    send NACK to $S_i$ with new $maxseqno_i$;
                else if $PTS_i + seqno_i$ equals $maxseqno_i$
                    of last probe then
                    // receiver is the bottleneck and
                    // now needs less additional packets
                    send NACK to $S_i$ with new
                    lower $maxseqno_i$;

        else // $TRPC \geq K$ or $t > D$
            leave multicast group;
            decode the received packets if $t \leq D$
```

(b) Receiver algorithm

Figure 3: Pseudo Code for the SomeCast-enabled Protocol

ns Prototype Implementation of SomeCast-enabled Protocol: We prototyped an implementation of our SomeCast-enabled protocol using the UCB/LBNL/VINT network simulator, ns-2.1b4 [30]. A new agent, called *scast*, is created as a subclass of AgentClass and defined in `scast.cc` and `scast.h`. This agent implements the SomeCast-enabled protocol. The primary sender starts transmitting data at time 25.0 (a warm-up period during which cross traffic at all links are generated). Other senders start transmitting as soon as one or more receivers join their multicast groups. The simulation run is stopped once *all* receivers receive by the deadline the needed packets to recover the original data, or whenever the simulation clock exceeds the deadline. In the latter case, one or more receivers had missed their deadline.

The details of our SomeCast-enabled protocol were described in Section 4.2. In our experiments, we take the total number of senders/groups to be 5. Each sender is assumed to have $2K$ encoded packets. We consider two variations of our protocol: (1) **SomeCast-1** where a receiver starts by joining the multicast group of primary sender S_0 and then joins other groups as needed; and (2) **SomeCast-5** where a receiver starts by joining all multicast groups of all 5 senders and then leaves groups as needed as long as it is able to receive by the deadline the number of packets it needs for full recovery.

ns Prototype Implementation of ManyCast and Unicast Protocols: To compare against SomeCast, we also simulated the two special cases of **ManyCast** and **UniCast**. ManyCast is the same as SomeCast except that joins and leaves are static, i.e. every receiver is subscribed to all multicast groups all the time. In Unicast, there is only one sender, the primary one, that every receiver joins.

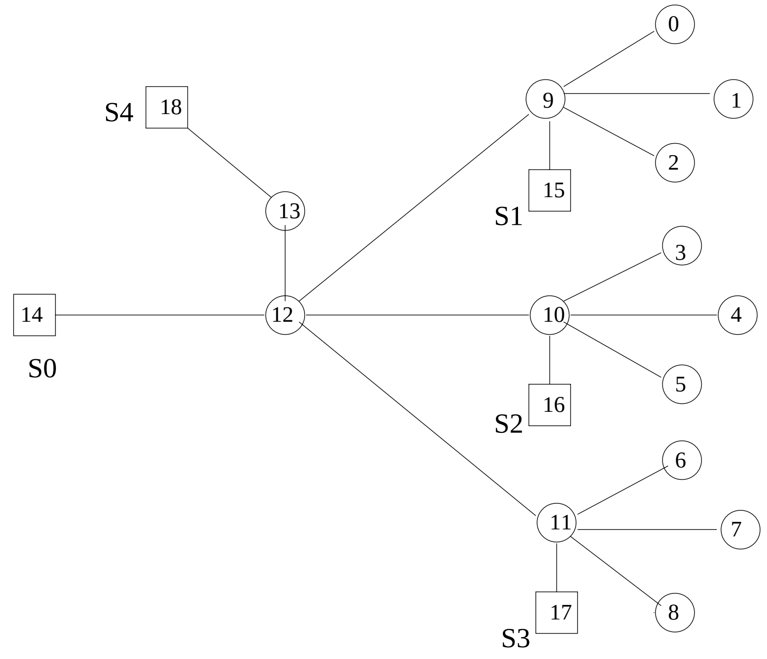

Figure 4: The network topology used in our simulations

5.2 Performance Measures

To compare our SomeCast protocol to the ManyCast and UniCast based protocols, we consider the following performance metrics:

Percentage Guaranteed: Percentage of receivers which successfully receive by the deadline (D) all packets (K) needed for full recovery.

Goodput: Ratio of the total number of (useful) packets received by D and used for full recovery at all receivers to the total number of packets sent by all senders.

Average Number of Groups Joined: Average number of multicast groups that a receiver joins.

In our simulations, we assume (for simplicity) that all receivers are subject to the same deadline. We define the *laxity* to be the ratio between the requested deadline and the most stringent deadline that can be met (i.e. when there are no losses and when every receiver joins the multicast groups of all 5 senders).

5.3 Simulation Results

We present our performance measures as a function of laxity. Figures 5 through 7 show Percentage Guaranteed, Goodput and Average Number of Groups Joined that were defined in Section 5.2.

Since in our SomeCast protocol, a receiver adjusts dynamically the number of groups to which it subscribes, SomeCast strikes a good balance between Percentage Guaranteed and Goodput since a receiver joins the *minimum* number of groups needed to receive K packets by the deadline. UniCast, where every receiver only joins the primary sender, yields the lowest Percentage Guaranteed, but the highest Goodput. Finally, ManyCast, where every receiver joins all 5 groups, yields the highest Percentage Guaranteed at the expense of lower Goodput. Observe that at lower laxities, ManyCast has the highest Goodput as it is able to make use of transmissions from all 5 senders to satisfy the requested deadline at many receivers. Other protocols suffer from lower Goodput as many receivers fail to meet their deadlines and transmissions, albeit from fewer senders, are wasted.

Figures 8 through 10 compare SomeCast-1 and SomeCast-5, where a receiver starts from 1 sender and 5 senders, respectively. Starting from 5 senders, a receiver in SomeCast-5 gradually unsubscribes from senders as long as it can receive K packets by the deadline. As expected, SomeCast-5 has a higher Percentage Guaranteed, i.e. it can meet more stringent deadlines than SomeCast-1. This is at the expense of lower Goodput as SomeCast-5 attempts to overcome not only losses due to cross traffic, but also losses due to interference among multiple senders over common links.

Figure 10 shows that for higher laxities, under SomeCast-5, receivers tend to join more multicast groups on average. This is because at first, a receiver unsubscribes from some of the 5 groups to which it initially subscribes, which is expected especially when deadlines are not tight. However, this increases the time taken to receive the K packets. This delay increases the chance of losses due to interference among multiple senders over common links, which in turn causes receivers to join more groups so as to meet their deadlines, thus resulting in degradation in goodput. Note

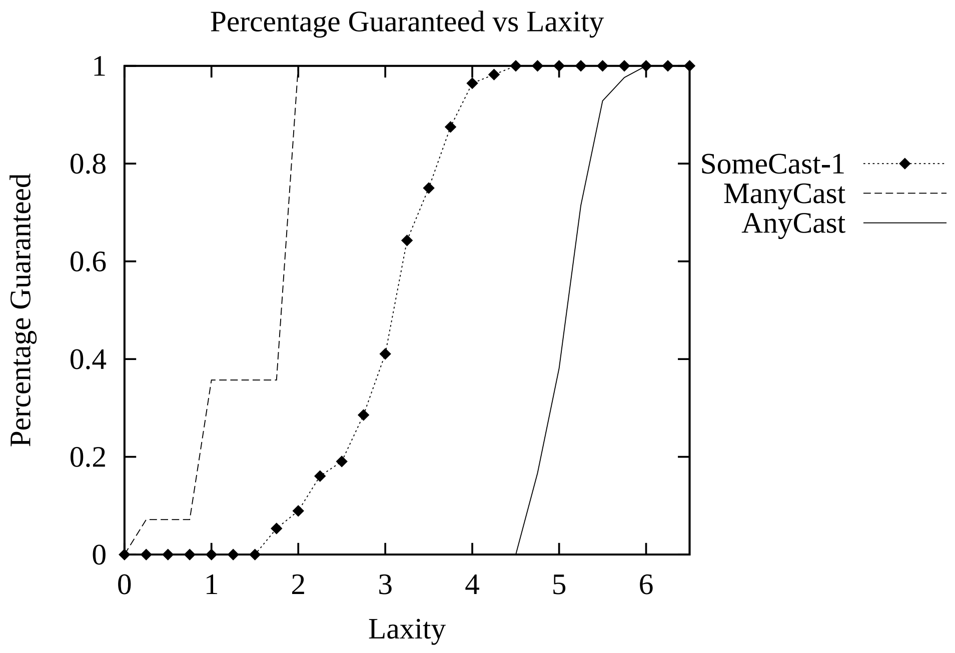

Figure 5: Percentage Guaranteed versus Laxity.

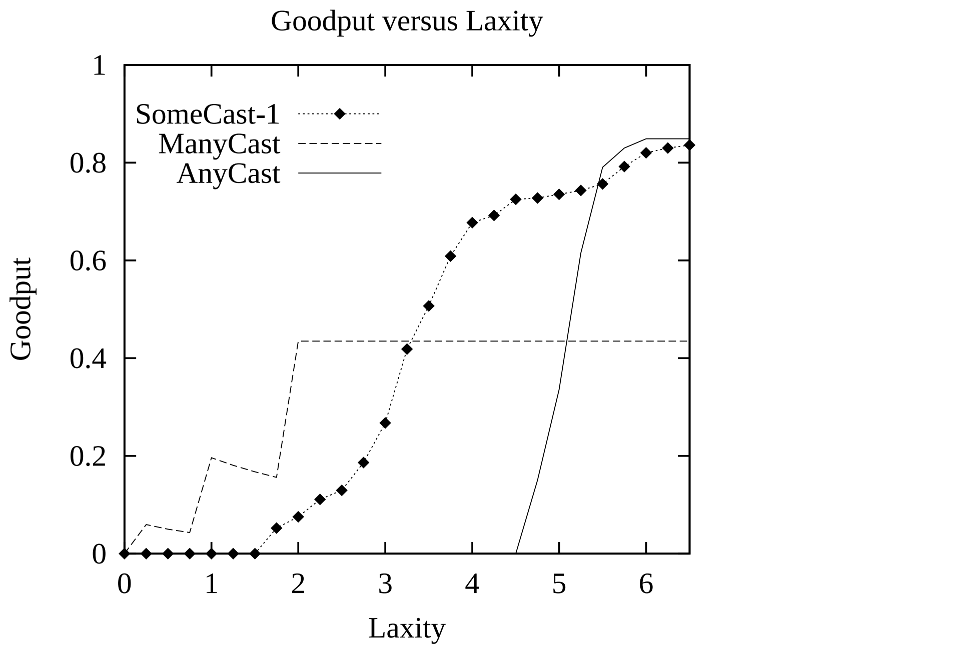

Figure 6: Goodput versus Laxity.

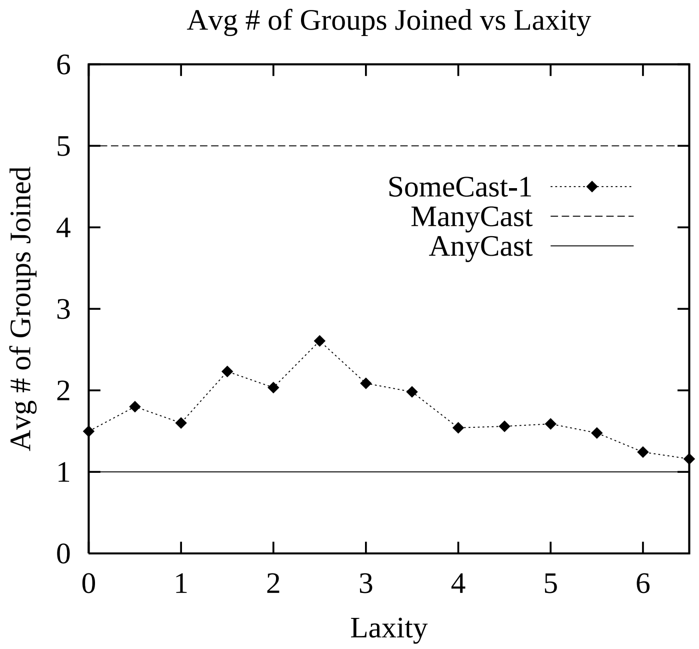

Figure 7: Average No of Groups versus Laxity.

that additional losses due to traffic from multiple senders can be reduced if the senders to which a receiver chooses to subscribe are chosen intelligently based on the tightness of the delay bound (laxity) and the loss rates on the paths from different senders. We will investigate this in a future paper.

6 Conclusion

Summary: We have proposed SomeCast—a novel paradigm for the reliable multicast of real-time data to a large set of receivers over the Internet. SomeCast enables receivers to adapt dynamically to the unpredictability of network conditions. This adaptation enables receivers to meet specific real-time QoS constraints. We have presented an instance of a SomeCast-enabled protocol, which we have prototyped under the UCB/LBNL/VINT network simulator (ns-2.1b4). We have presented simulation results that show the superiority of SomeCast when compared to the previously proposed ManyCast and AnyCast paradigms.

Future Work: In this paper we have not exploited the ability of receivers to *choose* the subset of multicast groups to which they subscribe. We are currently evaluating novel *multicast group selection* algorithms, which take into consideration knowledge of static network topology (e.g. based on distance between receiver and sender or closest router carrying the multicast group) or dynamic network conditions (e.g. whether or not paths to two multicast groups share a common congestion).

One of the salient features of the SomeCast paradigm is that it delegates the responsibility of QoS management to the receivers. As we indicated in this paper, such delegation is attractive

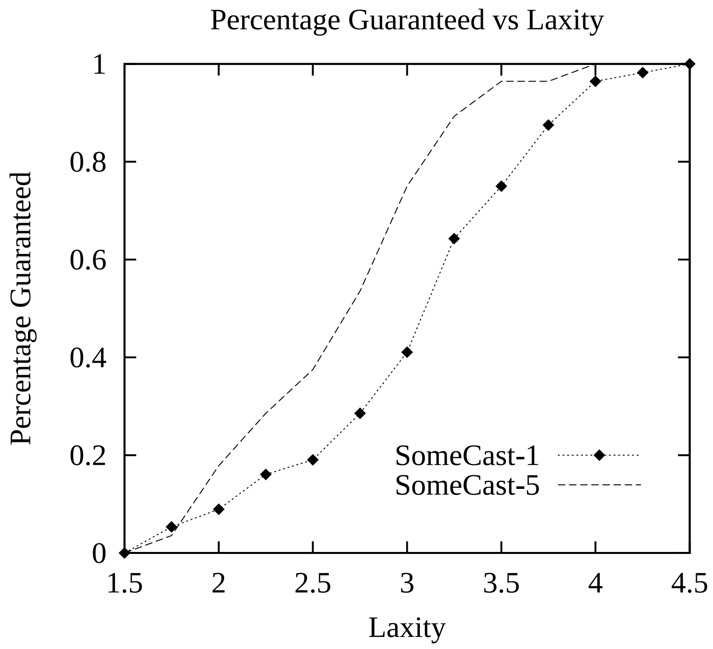

Figure 8: Guaranteed Ratio versus Laxity.

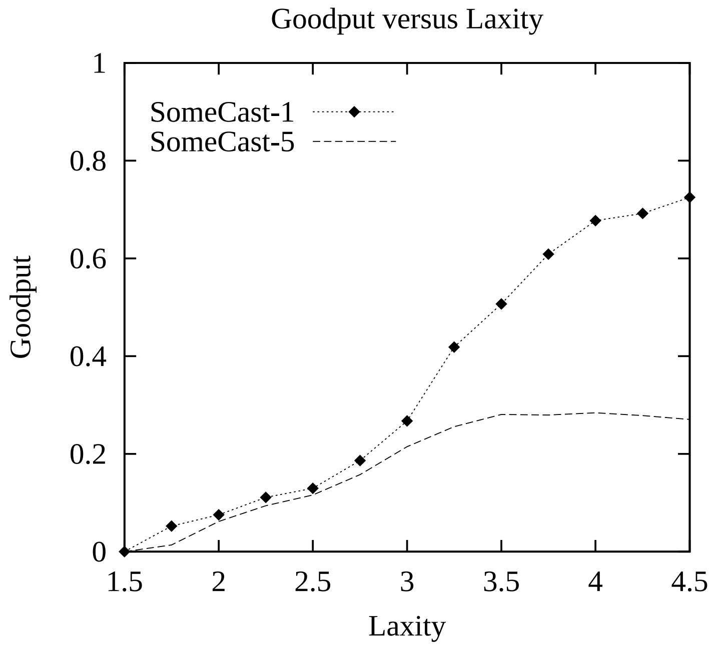

Figure 9: Goodput versus Laxity.

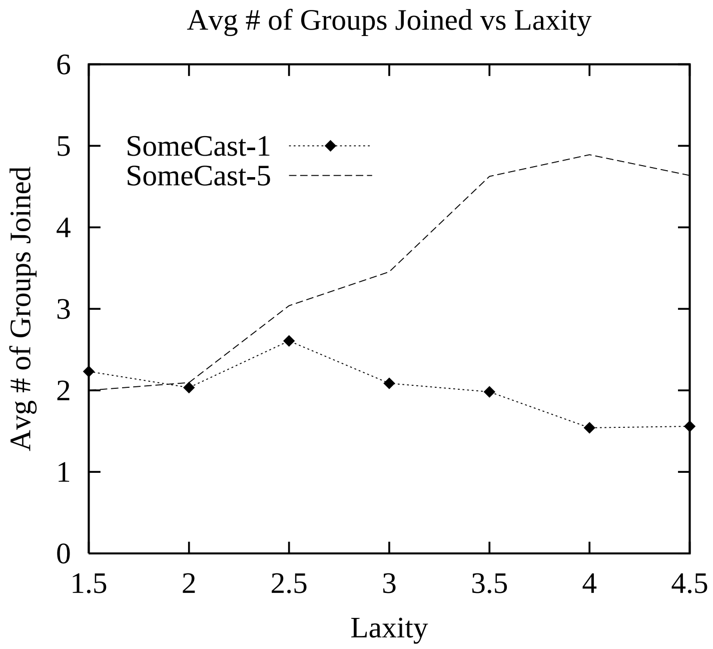

Figure 10: Average No of Groups versus Laxity.

because it boosts the scalability of the system by making the overhead incurred by senders *independent* of the number of receivers and/or the diverse characteristics of paths between senders and receivers. Another advantage of delagating the management of real-time QoS constraints to receivers is that it enables senders to respond to network congestion conditions without risking the violation of QoS constraints at the receivers (since receivers can recover from a reduction in the rate at which a sender transmits data on its multicast group). We are currently investigating such techniques, whereby SomeCast senders manage congestion by adopting TCP-friendly transmission policies (as opposed to the constant-bit-rate policy adopted in this paper).

Acknowledgments: We would like to thank John Byers for the many discussions on Tornado codes and Digital Fountains. This work was supported in part by NSF research grants ESS CCR-9706685, CAREER ANIR-9701988, and MRI EIA-9871022.

References

[1] Akamai Technologies. Freeflow Content Delivery System. http://www.akamai.com.

[2] K. Almeroth, M. Ammar, and Z. Fei. Scalable Delivery of Web Pages Using Cyclic Best-Effort (UDP) Multicast. In *Proceedings INFOCOM '98*.

[3] Azer Bestavros. AIDA-based Real-Time Fault-Tolerant Broadcast Disks. In *Proceedings of RTAS'96: The 1996 IEEE Real-Time Technology and Applications Symposium*, Boston, Massachusetts, May 1996.

[4] Sanjoy Baruah and Azer Bestavros. Real-Time Mutable Broadcast Disks. In Azer Bestavros and Victor Fay-Wolfe, editors, *Real-Time Database and Information Systems: Research Advances*, chapter 1, pages 3–22. Kluwer Academic Publishers, Norwell, Massachusetts, 1997.

[5] Azer Bestavros. An Adaptive Information Dispersal Algorithm for Time-critical Reliable Communication. In Ivan Frisch, Manu Malek, and Shivendra Panwar, editors, *Network Management and Control, Volume II*, chapter 6, pages 423–438. Plenum Publishing Corporation, New York, New York, 1994.

[6] Azer Bestavros and Gitae Kim. TCP Boston: A Fragmentation-tolerant TCP Protocol for ATM Networks. In *Proceedings of Infocom'97: The IEEE International Conference on Computer Communication*, Kobe, Japan, April 1997.

[7] J. Byers, Luby, and Mitzenmacher. A Digital Fountain Approach to Reliable Distribution of Bulk Data (Tornado). In *Proceedings of ACM SIGCOMM '98*, Vancouver, September 1998.

[8] John W. Byers, Michael Luby, and Michael Mitzenmacher. Accessing Multiple Mirror Sites in Parallel: Using Tornado Codes to Speed Up Downloads. In *Proceedings of IEEE INFOCOM '99*, pages 275–83, March 1999.

[9] Robert L. Carter and Mark E. Crovella. Dynamic Server Selection using Bandwidth Probing in Wide Area Networks. In *Proceedings of Infocom '97, the Sixteenth Annual Joint Conference of the IEEE Computer and Communication Societies*, April 1997.

[10] Mark Crovella and Azer Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, December 1997.

[11] S. Deering. Multicast Routing in a Datagram Internetwork. Tech. Rep. No. STAN-CS-92-1415, Stanford University, California, Dec. 1991.

[12] A. Feldmann, A. C. Gilbert, and W. Willinger. Data Networks as Cascades: Investigating the Multifractal Nature of Internet WAN traffic. In *Proceedings of SIGCOMM '98*, pages 42–55, October 1998.

[13] A. Feldmann, P. Huang, A. C. Gilbert, and W. Willinger. Dynamics of IP traffic: A Study of the Role of Variability and the Impact of Control. In *Proceedings of SIGCOMM '99*, September 1999.

[14] Z. Fei, S. Bhattacharjee, E. Zegura, and M. Ammar. A Novel Server Selection Technique for Improving the Response Time of a Replicated Service. In *Proceedings of INFOCOM '98*, San Francisco, CA, April 1998.

[15] S. Floyd, V. Jacobson, L. Ching-Gung, S. McCanne, and L. Zhang. A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing. In *Proceedings of ACM SIGCOMM '95*, pp. 342-356, Aug. 1995.

[16] M. Furini, D. Towsley. Real-Time Traffic Transmission over the Internet. Available as UMass CMPSCI Technical Report 99-73.

[17] C. Huitema. The Case for Packet Level FEC. In *Proceedings of IFIP 5th International Workshop on Protocols for High Speed Networks (PfHSN'96)*, France, October, 1995.

[18] Inktomi Corporation. Content Distributor. http://www.inktomi.com/products/network/cds/distributor.html.

[19] S. Kasera, G. Hjalmtysson, D. Towsley, J. Kurose. Scalable Reliable Multicast Using Multiple Multicast Channels. In *ACM SIGMETRICS '97*, Seattle, WA, June 1997.

[20] R. Kermode. Scoped Hybrid Automatic Repeat Request with Forward Error Correction (SHARQFEC). In *ACM SIGCOMM '98*, September 1998, Vancouver, Canada.

[21] Jia Rhu Li, Sungwon Ha, and Vaduvur Varghavan. Supporting heterogeneous packet flows in the Internet. BU/NSF Workshop on Internet Measurement Instrumentation and Characterization. Boston University, Boston, August 1999.

[22] John C. Lin and S. Paul. RMTP: A Reliable Multicast Transport Protocol. In *IEEE INFOCOM '96*, March 1996, pp. 1414-1424.

[23] M. Luby, M. Mitzenmacher, A. Shokrollahi, D. Spielman. Practical Loss Resilient Codes. In *Proceedings of the* 29^{th} *ACM Symposium on Theory of Computing*, 1997.

[24] M. Luby, M. Mitzenmacher, A. Shokrollahi. Analysis of Random Processes via And-Or Tree Evaluation. In *Proceedings of the* 9^{th} *Annual ACM-SIAM Symposium on Discrete Algorithms*, January, 1998.

[25] Yuh-Dauh Lyuu. Fast Fault-Tolerant Parallel Communication and On-Line Maintenance using Information Dispersal. Technical Report TR-19-1989, Harvard University, Cambridge, Massachusetts, October, 1989.

[26] Jamshid Mahdavi and Sally Floyd. The TCP Friendly Web Site. Technical note sent to the end2end-interest mailing list, January 8, 1997.

[27] A. McAuley. Reliable Broadband Communication Using a Burst Erasure Correcting Code. In *ACM SIGCOMM '90*, Sep. 1990, Philadelphia, pp. 297-306.

[28] D. Mills. Network Time Protocol (version 3). Request For Comments, RFC 1305, March 1992.

[29] J. Nonnenmacher, E. Biersack, D. Towsley. Parity-Based Loss Recovery for Reliable Multicast Transmission. In *Computer Communications Review ACM SIGCOMM*, vol. 27, No. 4, 1997.

[30] UCB/LBNL/VINT Network Simulator, ns, URL: *http://www-mash.cs.brekeley.edu/ns.*

[31] Padhye, J. Kurose, D. Towsley, R. Koodli. A Model Based TCP-Friendly Rate Control Protocol. Proc. *IEEE NOSSDAV'99* (Basking Ridge, NJ, June 1999). An Extended Abstract of an earlier version appeared in Proc. ACM SIGMETRICS'99 (Atlanta, GA, May 1999).

[32] Kihong Park, Gitae Kim, and Mark E. Crovella. On the Effect of Traffic Self-Similarity on Network Performance. In *Proceedings of SPIE International Conference on Performance and Control of Network Systems*, November 1997.

[33] Request For Comments, RFC 2117. Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification. June 1997.

[34] Michael O. Rabin. Efficient Dispersal of Information for Security, Load Balancing and Fault Tolerance. *Journal of the Association for Computing Machinery*, 36(2):335–348, April 1989.

[35] Irving S. Reed and Gustave Solomon. Polynomial Codes over Certain Finite Fields. *Journal of the Society for Industrial and Applied Mathematics*, 1960.

[36] L. Rizzo. Effective Erasure Codes for Reliable Computer Communication Protocols. In *ACM Computer Communications Review*, vol. 27, n.2, Apr. 1997, pp. 24-36.

[37] L. Rizzo and L. Vicisano. A Reliable Multicast data Distribution Protocol based on software FEC techniques. In *Proceedings of the Fourth IEEE, HPCS'97 Workshop*, Chalkidiki, Greece, June 1997.

[38] D. Rubenstein, J. Kurose, D. Towsley. Real-Time Reliable Multicast Using Proactive Forward Error Correction. In *NOSSDAV '98*, Cambridge, UK, July, 1998.

[39] Digital Island (formerly known as SandPiper Networks). Sandpiper's Content Distribution Network. http://www.sandpiper.com.

[40] StarBurst Software. One-To-Many Content Distribution in an Enterprise Environment. A white paper. Available from http://www.strburst.com.

[41] W. Strayer, B. Dempsey, and A. Weaver. *XTP: The Xpress Transfer Protocol.* Addison-Wesley, URL: *http://hegschool.aw.com/cseng/authors/dempsey/xtp/xtp.nclk.*

[42] R. Yavatkar, J. Griffioen, and M. Sudan. A Reliable Dissemination Protocol for Interactive Collaborative Applications. In *Proceedings of the ACM Multimedia '95 Conference*, November 1995.