A Utility-Aware Middleware Architecture for Decentralized Group Communication Applications

Jianjun Zhang[1], Ling Liu[1], Lakshmish Ramaswamy[2], Gong Zhang[1], and Calton Pu[1]

[1] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
{zhangjj,lingliu,gzhang3,calton}@cc.gatech.edu
[2] Department of Computer Science, University of Georgia, Athens, GA 30602, USA
laks@cs.uga.edu

Abstract. Although unstructured Peer-to-Peer (P2P) networks provide economical platforms for supporting group communication applications, their ad-hoc nature poses significant challenges to the performance of the group communication services. This paper presents the design and implementation of GroupCast − a utility-aware middleware architecture for scalable and efficient P2P group communications. The GroupCast design is characterized by two unique features. First, we present a *utility function* for quantifying the role of unicast links in enhancing the scalability and efficiency of the group communication applications. Second, we develop a utility-aware distributed spanning tree construction algorithm for efficiently propagating group communication messages. It dynamically creates and maintains the group communication channels by optimizing the utility value of the group communication spanning trees. In addition, we also outline a utility-based overlay management protocol for constructing and maintaining low-diameter overlay networks. Our experiments show that the GroupCast system can improve the scalability of wide-area group communication services by one to two orders of magnitude.

Keywords: Peer-to-peer systems, Overlay networks, Utility functions.

1 Introduction

Multi-party group communication applications such as multi-player online games, online community based advertising, real-time conferencing [3], and instant messaging [2] have experienced a surge of popularity in the past few years. The applications are characterized by exchanges of textual or multimedia contents among multiple participants. Decentralized Peer-to-Peer (P2P) networks have evolved as a promising paradigm for providing open and distributed information sharing services by harnessing widely distributed, loosely coupled, and inherently unreliable computer nodes (peers) at the edge of the Internet. The success of Skype [5] has demonstrated both the opportunity and the feasibility of utilizing P2P networks as economical infrastructures for providing wide-area group

R. Cerqueira and R.H. Campbell (Eds.): Middleware 2007, LNCS 4834, pp. 410–428, 2007.
© IFIP International Federation for Information Processing 2007

communication services. However, the overlay networks in Skype are used only for service lookup and control signaling. Under the multi-party conference settings, each node is required to send the payloads directly to other participants of the communication group through its IP unicast links [9]. This places severe limitations on the scalability of multi-party conference calls in Skype.

The natural questions that come up include: *Can P2P overlays be utilized for implementing scalable group communication services over wide-area networks? If so what techniques and system level optimization are critical for enhancing the efficiency and scalability of decentralized wide area group communications?* Although several researchers have explored a related problem in the context of designing application-level multicasting or end-system multicasting (ESM) schemes [20,22] over P2P overlays, surprisingly most of these works are designed to work in conjunction with structured P2P networks, and they rely on the distributed hash table (DHT) abstractions of the P2P network for inter-peer communication and routing [11,21]. However, it is widely recognized that in environments that exhibit high *churn* rates maintaining DHT-based structures imposes severe overheads, which can affect the performance of the applications running on top of these networks to a considerable extent [13]. In contrast, *unstructured P2P networks* like Gnutella [25] are simple to implement, have low maintenance costs, and provide better resilience to network churn caused by peer entries, exits, and failures. To the best of our knowledge, very few group communication applications have been implemented on top of unstructured P2P networks. We hypothesize that common concerns about the non-deterministic nature of communication and service lookup in unstructured overlay networks and their inefficient utilization of the underlying IP network resources are the main reasons for the lack of work in this area.

Designing scalable group communication services on top of unstructured P2P networks poses three main challenges. The first challenge is to translate wide-area group communication application requirements such as communication efficiency and system scalability into the metrics that can be used while designing the communication structures and managing the topologies of the overlay networks. Second, the unstructured P2P networks suffer from heavy messaging overheads and high service lookup latencies. The challenge is to devise low-cost service lookup mechanisms that are effective for both control signaling and communication group management. The resilience of unstructured P2P overlays to network churn is rooted in the fact that they do not use any global control mechanisms for regulating resource distribution and the network topology. The third challenge is to design overlay network management protocols such that the features critical to the performance of group communication applications are incorporated without trading away the inherent randomness of these overlays.

Towards addressing these challenges, this paper presents *GroupCast* - a utility aware decentralized middleware architecture for scalable and efficient P2P group communication applications. In designing the GroupCast system, this paper makes two unique contributions:

- First, we propose a utility function to quantify the usefulness of unicast links to the efficiency of individual communication groups as well as to the scalability of the entire group communication infrastructure. This utility function provides a careful combination of the two most important factors that influence the performance of the system, namely *network proximities of the peers* and *resource availabilities at the end hosts*.
- Second, we design a *utility-aware* mechanism for constructing spanning trees required for disseminating group communication payloads. The objective of this scheme is to optimize the utility-values of the resultant spanning trees. Further, considering the decentralized nature of unstructured P2P networks, this scheme has been designed to operate in a completely distributed fashion, and it does not rely upon any global topological information.

In addition to the above contributions, we also outline a *utility-based* P2P overlay network management protocol that uses the proposed generic utility function for constructing low-diameter unstructured P2P overlays that are comparable to structured P2P network in their scalability and efficiency. This paper presents several experiments to evaluate the utility-aware middleware architecture and its component techniques. The results show that the proposed techniques provide significant scalability and efficiency benefits for the group communication applications.

2 The Basic P2P-Based Group Communication Framework

Spanning tree forms the fundamental structure in most group communication schemes. The spanning tree is an acyclic overlay connecting all the participants of a communication group. The group communication messages (payload) are disseminated through the spanning tree so that they reach all the participants. The various group communication schemes differ in the manner in which they construct and maintain the spanning trees. Our system employs multi-level spanning trees for achieving the scalability needed for supporting group communication in large wide-area networks. The proposed framework includes completely distributed strategies for building and maintaining spanning trees of communication groups.

We introduce a few notations that would be used in the rest of the paper. The P2P network is conceptualized as a directed graph $G < V,\ E >$, where $V = \{p_0, p_1, p_2, \ldots, p_{N-1}\}$ represents the peers in the network and $E = \{e_0, e_1, \ldots e_{M-1}\}$ denotes the logical links in the network. The spanning tree $T_Pt < V_{Pt},\ E_{Pt} >$ is defined as a connected, acyclic sub-graph of G, where the participant set $V_{Pt} \subseteq V$ and links set $E_{Pt} \subseteq E$. Each peer is aware of only its immediate neighbors. Further, the network does not have any distributed hash table (DHT) abstractions.

2.1 Constructing a Distributed Spanning Tree

One of the challenges in developing group communication systems is to design a completely distributed scheme for building spanning tree. Several application level multicast (end-system multicast) systems have addressed a very similar problem (the problem of constructing multicast trees). However, as we explain below, none of them are directly applicable for building spanning trees on unstructured P2P networks.

The existing multicast tree construction schemes can be classified into three broad categories. In the first approach, the participants of a multicast group explicitly choose their parents in the multicast tree from a list of candidate nodes [7,18,17]. Due to the complexity of those protocols, there are very few actual implementations of these algorithms. The second approach, which is adopted by systems like Narada [14] and Scattercast [12], constructs the spanning tree in two-steps. The first step constructs a well-connected mesh from the nodes in the network. The second step uses this mesh structure and constructs shortest path spanning trees through well-known distributed algorithms. However, these systems do not scale well, especially when the underlying network experiences considerable churn. The third approach, represented by systems like CAN-multicast [21] and SCRIBE [11], assumes that the nodes of the underlying network are organized as a structured P2P network [20,22]. The multicast tree is constructed using the deterministic routing functionalities of these P2P networks. As we discussed in Section 1, DHT-based structured P2P networks are not suitable for scenarios wherein the peer populations are transient. In short, none of the current multicast tree construction approaches are applicable for the problem at hand.

We have developed completely decentralized scheme for building group communication spanning trees on a generic unstructured P2P network. We leverage techniques such as selective message forwarding for reducing the communication costs of spanning tree construction and maintenance.

2.2 Building the Communication Group

The objective of our communication group construction algorithm is to select the edges or the links in the P2P overlay to form the spanning tree that connects all the group participants. The communication group construction algorithms usually includes the implementation of two functionalities. First, participants should be aware of the existence of the communication group to which they will join. Second, a newly joined participant should be able to setup a connection to the existing nodes in the chosen communication group for sending and receiving the communication payloads.

The first task is usually accomplished by appointing a node as the rendezvous point or the multicast source, and publishing the node's information at a well-known location such as a bulletin board system. Two strategies have been proposed for implementing the second functionality. The first scheme is similar to the DVMRP IP-multicast protocol [16]. Instead of using the IP level network devices such as routers to implement the polling and pruning processes of multicast

group management, this strategy uses overlay networks and peers. This strategy is adopted by the Scattercast system [12], in which the source node solely advertises route information and each node in the overlay forwards this advertisement and builds the local routing table entries. To remove loops and to avoid the problem of counting-to-infinity, the full path information is embedded into the forwarded advertisement messages. We refer to this scheme as *Non-Selective Service Announcement*(NSSA) scheme. In the second strategy, the multicast source is mapped to a well-known node serving as the rendezvous point, and the subscribers use this identifier as the keyword in their subscribing requests [11]. The structured system topology and the deterministic routing algorithms decide the series of peers through which each subscription request would be forwarded so that it reaches the rendezvous point or an existing participant in the multicast group. The reverse of this path would be used for forwarding the multicast payloads down from the multicast source.

Two characteristics of our system prevent us from directly reusing these schemes. First, the nature of group communication applications is different from end-system multicast systems. In end-system multicast systems, communication payloads are forwarded in one direction in most of the cases (from the multicast source to all the other nodes), while in group communication systems, each participant may initiate messages in addition to receiving them. Second, the unstructured nature of our P2P overlay prevents us from directly using the reverse of the searching path as the payload communication path.

We have proposed a scheme that combines the advantages of these two schemes, while avoiding their disadvantages. We call our scheme the *Selective Service Announcement* (SSA) scheme. In this scheme, the spanning tree for a communication group is established in three steps.

Step 1 - Choosing Rendezvous Point: First, a peer in the P2P overlay is chosen as the rendezvous point. Unlike the rendezvous point in SCRIBE [11], to which all the multicast payloads are first forwarded, our rendezvous point serves as the source of the group advertisement messages and will behave as a normal node in the communication spanning tree. There are several ways to choose such a rendezvous point. It can be setup as a dedicated server donated by a service provider who injects contents into the communication group. For groups that are setup for applications like online conferences, the first participant can initiate a random walk search to locate a node that has enough access network bandwidth and computational power to act as a rendezvous point.

Step 2 - Advertising: In the second phase, the rendezvous point advertises the group information to the potential participants of the communication group. The flooding scheme used for similar purposes in DVMRP [16] and Scattercast [12] incurs redundant messages in the overlay network. Our SSA scheme alleviates the communication overheads in the following manner. In our scheme, each peer that receives the advertisement message will forward it to a few of its neighbors, rather than flooding the message to all neighboring nodes. Our basic group communication framework uses a very simple approach for selecting neighbors,

namely the random strategy. In this algorithm, the rendezvous point and the other nodes receiving the advertisement message randomly select a pre-specified fraction of their neighbors and send them the message. The message propagation terminates when the TTL becomes zero. However, this simple advertisement scheme suffers from two major drawbacks. We discuss these limitations later in the paper and present schemes for mitigating them.

Step 3 - Subscribe: Subscription activities are initiated when a peer p_i decides to join a communication group. Two scenarios need to be considered. First, if the potential service subscriber (peer p_i) has already received and routed the service advertisement, then it is already on the message forwarding path of this communication group. All it needs to do is to start the subscription process by sending the joining message in the reverse direction of incoming SSA message. However, note that the advertisement message might not reach all potential subscribers. In case the subscriber has never received the SSA message, a search method provided by the P2P overlay is triggered to look up the neighborhood of the peer for discovering nodes that might have received the SSA advertisement message.

The search method is implemented as a ripple search in standard Gnutella P2P network, with initial TTL (Time to Live) value set to a very low value. Because our advertisement mechanism would have already *pushed* the service information to different topological regions of the network, a potential subscriber can find a nearby neighbor that has received the SSA message with high probability. Our experiment reports that the average success rate of subscription search is close to 100%, even when the TTL of the search messages are set to 2. Once such a node is discovered, the subscription message is sent to it, which then forwards in the reverse direction of the original SSA message.

2.3 Limitations of the Basic Framework

The basic group communication framework has two important limitations which can affect its efficiency and scalability. The first limitation is the manifestation of the overlay-underlay mismatch problem. Since, in the advertisement phase of the scheme, a node receiving the advertisement forwards the message to a *randomly* chosen subset of neighbors, the resulting tree might not always be efficient in terms of the relative locations of its nodes on the physical network. For example, a node p_i located in New York might have a node p_j located in Australia as one of its children, which in-turn might have a child p_l located in Boston. This has a negative effect on the latencies experienced by the group communication messages. Similarly, the capability (resource availability) of a node p_i in the spanning tree might be completely different from the capabilities of its parents or children. This mismatch among the capacities of the neighbors in the spanning tree can result in high packet losses. This again affects the performance of group communication.

We propose two middleware level techniques for overcoming the above drawbacks, namely a utility-aware spanning tree construction scheme and a

utility-aware topology management scheme for the underlying P2P network. While the first technique addresses the question as to *how should the connections in the overlay be utilized for group communication applications?*, the second technique addresses the question of *how the peers should choose and maintain their neighbors in the overlay?* However, it is interesting to observe that these two questions are the manifestations of the same design issue, namely *given a list of nodes, say L, what are the metric(s) that dictate which of these nodes a peer p_i should connect to?* Both these techniques rely upon a unique utility-function, which assigns different preferences (rankings) to each peer in the list L. In the next section, we explain the formulation of the utility function. We then describe how this utility function is utilized in the proposed techniques.

3 The Utility-Aware Middleware for P2P Group Communication

This section focuses on the two main components of the GroupCast design. First, we describe the utility function we use to quantitatively model the critical performance metrics of wide area group communication applications. Second, we discuss how to employ our utility function to optimize the group communication channel construction and maintenance by developing a utility-aware distributed spanning tree construction algorithm that can efficiently propagate group communication messages. Finally, we also outline our utility-based overlay management protocol which provides the capability for constructing and maintaining low-diameter overlay networks to further enhance the performance of the group communication services.

3.1 The Utility Function

The group communication in overlay networks essentially occurs by forwarding the communication payload through unicast IP network links. Hence, the properties of the unicast links interconnecting peers in the P2P overlay largely decide the performance and the efficiency of the group communication system. Our utility function considers the two important factors that determine the performance of unicast links, namely the network proximity of the end-nodes and the similarity between among the capacities of the peers. The network proximity between the end-hosts determines the latency of the unicast link. Similarly, it is known that mismatch between the packet-forwarding workloads and the capacities of peers introduces bottlenecks in the communication overlay and may result in high packet losses. We note that these two factors might sometimes be counteracting. For instance, a peer in the list L which is closest to p_i, might have completely different resource availabilities than p_i. Our utility function provides a careful combination of these two factors based on the utility preference of peer p_i, as well as the desired performance properties of the entire overlay.

Concretely, for each node p_j in the list L (recall that L represents a list on potential nodes from which the peer p_i chooses a subset), we assume that two

types of information are available: the node capacity C_j, and the relative distance between peer p_i and peer p_j, denoted by $D(p_i, p_j)$. The capacity of a peer is measured in terms of its accessible network bandwidth, since the performance of a peer in a distributed environment like P2P networks is largely decided by its access network bandwidth available for forwarding communication payloads. The access network bandwidth can be specified by the end user in terms of the number of 64kbps connections the node is willing to support. Alternatively, it can also be estimated by network probing techniques. We use the network coordinates to estimate the relative distance between any two peers. Vivaldi [15] and GNP [1] are some of the techniques proposed for measuring the network coordinates of nodes in wide area networks.

We define two utility-based preference metrics based on the two important performance factors namely network proximity and node capacity. Given a list of peers L, we define the *Distance Preference* of peer p_i to peer $p_j \in L$ as the probability that peer p_i chooses peer p_j out of L, based on the network coordinate distance between them. The closer the peer p_j is to peer p_i, the more likely it is chosen. The *Distance Preference* is computed as indicated in Equation 1.

$$PD_{p_i}(L, p_j) = \frac{\frac{1}{d_{p_i}(L,p_j)} - \alpha}{\sum_{p_k \in L} \frac{1}{p_{d_i}(L,p_k)} - \alpha} \tag{1}$$

where $\alpha \in (-\infty, 1)$ is a tunable parameter that indicates the degree to which p_i's prefers closer peers. Higher values of α indicates that p_i strongly prefers closer peers and vice-versa. We choose $\alpha < 1$ so that there is nonzero preference on each $p_j \in L$. The function $d_{p_i}(L, p_j)$ gives the *normalized distance* between p_i and p_j. $d_{p_i}(L, p_j)$ is defined as follows:

$$d_i(L, j) = \frac{D(p_i, p_j)}{\text{MAX}_{p_k \in L} D(P_i, p_k)} \tag{2}$$

Note that $0 < d_{p_i}(L, p_k) \leq 1$ for each peer p_k in the list L.

Similarly, we define the *Capacity Preference* utility metric of peer p_i with respect to peer p_j as the probability that peer p_i chooses peer p_j out of L based on the node capacity of peer p_j. The goal is to utilize higher capacity nodes to relay group communication messages to larger number of peers. Equation 3 gives the formulation for the *Capacity Preference* utility metric.

$$PC_{p_i}(L, p_j) = \frac{C_{p_j} - \beta}{\sum_{p_k \in L} C_{p_k} - \beta} \tag{3}$$

Here C_{p_j} is the node capacity of the peer p_j. The parameter $\beta \in (-\infty, 1)$ plays a similar role as that of α in equation 1.

While the *Capacity Preference* and *Distance Preference* encapsulate the utility of nodes in L from two different perspectives, we need a means to combine these two utility parameters into a single utility function. In this regard, it is interesting to observe that the peer p_i which wants to select a subset of peers from L should also consider its own resource availability (capacity) while

making its choices. If the peer p_i possesses more resources, we would like to use it as a forwarding hub in the overlay network and applications. Such a peer should be connected to those peers that have similar resources and play similar roles in the overlay network, which would make it a member of the "core" of the overlay network. On the contrary, if the resources of peer p_i are limited, it should not be placed into the core as that would easily exhaust its resources. A better choice for such a limited resource peer would be to connect to peers that are physically closer to it and use them to access the overlay network. Hence, the weightage given to the two utility metrics (*Capacity Preference* and *Distance Preference*) depends upon the capacity of peer p_i. Accordingly, we define the combined utility function *Selection Preference* of peer p_i to peer $p_j \in L$ as a weighted combination of *Capacity Preference* and *Distance Preference*.

$$P_{p_i}(L, p_j) = \gamma \cdot PC_{p_i}(L, p_j) + (1 - \gamma) \cdot PD_{p_i}(L, p_j) \tag{4}$$

Here γ is the weightage factor such that $0 \leq \gamma \leq 1$.

The configurable parameters α, β, and γ gives us the flexibility to fine-tune the utility function for different application scenarios. For instance, in an overlay network supporting applications that are sensitive to network proximity, α can be set to higher values and γ to be set to lower values. This would ensure that network proximity is the dominating factor when peers make their choices. On the contrary, for an overlay network that emphasizes more on load balancing, a higher value for β and a higher value for γ would be more preferable.

The values of parameter α, β, and γ can be mathematically derived by using techniques similar to the ones used by Bu and Towlsey [10]. However, these techniques require information about the exact number of peers and the exact distributions of the various system-level parameters. In decentralized environments like P2P networks where global statistical mechanisms are expensive to implement, it is unlikely that such information would be available. The GroupCast system adopts an approximation approach to address this problem. Specifically, we define *Resource Level* r_i as the fraction of peers that have less capacity than peer p_i in the overlay network. r_i can be estimated by sampling a few peers that are known to p_i. We use the resource levels of various peers to set the three parameters as $\alpha = 1 - r_i$, $\beta = r_i$, and $\gamma = r_i^{-ln(r_i)}$. Substituting the values for α, β, γ, PC, and PD into equation 4, we obtain:

$$P_i(L, j) = r_i^{-\ln(r_i)} \cdot \frac{C_j - r_i}{\sum_{k \in L} C_j - r_i} + (1 - r_i^{-\ln(r_i)}) \cdot \frac{\frac{1}{d_i(L,j)} - (1 - r_i)}{\sum_{k \in L} \frac{1}{d_i(L,k)} - (1 - r_i)} \tag{5}$$

We note that this configuration reflects our design rationale. The β and γ parameters assume higher values for peers with higher capacities. Hence, these peers would give preference more powerful peers while choosing a subset from L. In contrast, for peers with less resources α assumes higher values whereas β and γ become small. Thus, for these peers the subset selection is predominantly based upon the network proximities. In other words, the less powerful peers connect to nodes that are closer to them. Further, they avoid peers with large capacities, thereby shielding themselves from getting overloaded.

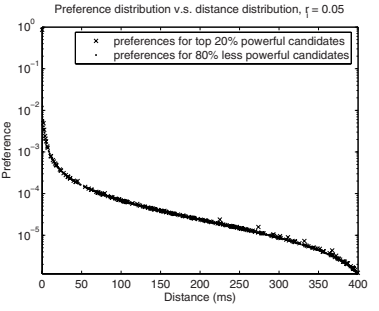

Fig. 1. Selection preference of low capacity peer vs. distance to other peers

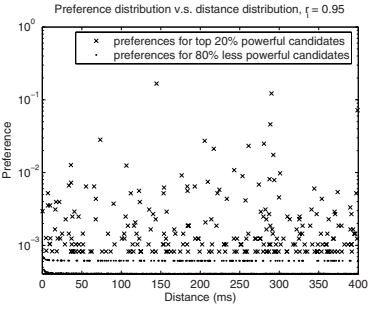

Fig. 2. Selection preference of high capacity peer vs. distance to other peers

To evaluate the effectiveness of the selection preference metric, we simulate the selecting process of three peers, using a set of synthetic data. We assign each of them with different resource level value. The one with $r_i = 0.05$ represents a peer with low capacity. Similarly, the one with $r_i = 0.5$ simulates a peer with medium capacity, and the one with $r_i = 0.95$ represents a powerful peer. For each of them, we generated a list of 1×10^3 peers, each of which is assigned a capacity value that follows a zipf distribution with parameter 2.0. We assume that the distance between each candidate peer and the peer evaluating them follows a uniform distribution Unif(0ms, 400ms).

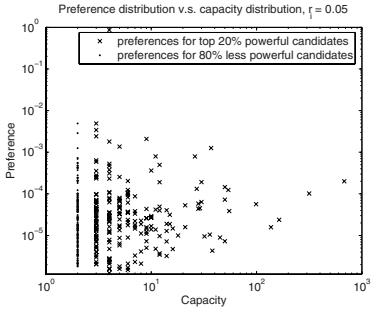

Fig. 3. Selection preference of low capacity peer vs. capacity of other peers

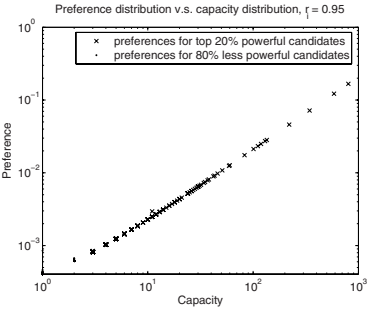

Fig. 4. Selection preference of high capacity peer vs. capacity of other peers

Figure 1 $\sim$ Figure 4 plot the simulation results, which exactly reflects our design rationale. For a weaker peer that has $r_i = 0.05$, the selection preference to other peers is dominantly decided by its distance to them, as plotted in Figure 1 and Figure 3. On the contrary, the selection preference of a powerful peer is largely decided by the node capacity of peers in the candidate set as shown in Figure 2 and Figure 4. For peers that has medium amount of resources, it equally prefers powerful and nearby peers [27].

3.2 Utility-Aware Spanning Tree Construction

In this section, we describe our technique for infusing utility-awareness into spanning tree construction for group communication. The central idea of this technique is to ensure that the edges in the spanning trees have very high utility values, thereby optimizing the overall group communication performance. If the topology of the P2P network and the utility values of all the unicast links in the network were to be available in a centralized location, we could have used one of the several optimization techniques for constructing utility-aware spanning tree. Unfortunately, due to the very nature of P2P systems collecting topological and utility information at a centralized location would be extremely expensive, if not impossible. Therefore, the challenge is to design a completely distributed spanning tree construction technique that is not only effective in ensuring high utility values for the edges in the tree but is also efficient and lightweight.

We observe that the basic spanning tree construction technique that we explained in Section 2 is indeed completely distributed, and it does not rely upon any centralized topological information. Therefore, the question is *whether it is possible to achieve high utility values while retaining the overall spanning tree construction framework?*

Our utility-aware spanning tree construction scheme is based upon the following crucial observation. Of the three phases of the basic spanning tree construction scheme, the advertisement phase has the most significant influence on the structure of the resultant spanning tree. In other words, the advertisement decisions made by various peers more or less determine the structure of the spanning tree. This is because, if a node p_l receiving an advertisement decides to participate in the group being advertised, the very links through which the advertisement was propagated to p_l from the rendezvous node would become a part of the corresponding spanning tree. However, in the basic group communication framework, each peer receiving the advertisement sends it to a randomly selected subset of its neighbors.

From the above observation, we conclude that the most natural way for injecting utility-awareness into the spanning tree construction process is to incorporate it at the advertisement phase. Accordingly, in the utility-based spanning tree construction technique, peer receiving the advertisement forwards it to a subset of its neighbors based on their utility values. Specifically, the probability of a neighbor being included in the subset selected for forwarding the advertisement is directly proportional to its utility value. Thus, a neighbor that has a higher utility value has a higher chance of being included in the subset of nodes to which the advertisement is forwarded.

Specifically, a rendezvous point rp evaluates the utility value of its neighbors using Equation 5. Based on these utility values, it chooses the peers either have similar capacities as rp or are physically close to rp, depending on the capacity of rp. These peers are the ones that are more useful to rp. They receive the advertisement and are likely to be included in the spanning tree.

Upon receiving an SSA message, an arbitrary peer p_k performs two tasks. First, peer p_k uses a local hashing table to check and record if it has already

received the same message from any other neighbors. The message will be dropped if it is a duplicated one. Otherwise, it uses a similar mechanism as that of the rendezvous point to select neighbors for further propagating the SSA messages.

In effect, when a peer receives an advertisement, it is more likely that the advertisement traversed a path in which each link had a high utility value. If this peer decides to participate in the group being advertised, the path of the advertisement becomes a part of the corresponding multicast tree. Thus, our scheme seamlessly incorporates utility awareness into the spanning tree construction process.

3.3 Utility-Aware Topology Management

Our utility-aware spanning tree construction algorithm builds the spanning tree from existing connections of the overlay. Thus, the performance of the resultant spanning trees depend upon the topology of the underlying P2P network. With the aim of further enhancing the performance of the GroupCast middleware, we have designed a utility-aware overlay construction mechanism. In this section, we briefly outline the mechanism.

The objective of the utility-aware overlay construction technique is to create P2P networks in which the neighbors of an arbitrary node p_i have reasonably high utility values with respect to p_i. Unlike many P2P networks that are based on the concept of super nodes, our technique inserts both high-capacity and low-capacity peers into the same overlay. Our technique essentially works as follows: When a peer p_i joins the overlay, it gathers the information of a number of existing peers as its neighbor candidates. The new peer calculates the probability of connecting to each candidate by using the utility function defined in Equation 5. These probabilities and the total number of connections that the p_i intends to maintain determine whether p_i would establish a connection with an arbitrary neighbor candidate peer.

Specifically, a joining peer p_i obtains a list of existing peers either using its local cache which contains its P2P network neighbors carried from the last session of activities or by contacting a host cache server. Upon receiving a query request from peer p_i, the host cache sorts its cached entries in the ascending order by their network coordinate distances to peer p_i. From the top of this sorted list, the host cache selects a list of peers BD_i. They are returned to peer p_i together with a list of randomly selected peers BR_i. Starting from the subset B_i of bootstrapping peers received upon its entry, Peer p_i sends a probing message M_{prob} to each peer $p_k \in B_i$. Each peer p_k that receives this probing message sends back a responding message M_{prob_resp}, which is augmented with a list of p_k's P2P network neighbors $Nbr(p_k)$. Peer p_i assembles all the neighbor information contained in the probing replies and compiles them into a candidate list LC_i. For each unique peer $p_j \in LC_i$, peer p_i computes two types of information: (1) The *occurrence frequency* of peer p_j, which records the number of appearances of peer p_j in LC_i, denoted as $f_i(p_j)$. As LC_i serves as a sampling of the peers in the network, $f_i(p_j)$ is the sample of the degree of peer p_j. (2) The estimation of the

physical network distance between peer p_i and peer p_j, denoted by $d_i(LC_i, p_j)$, as defined in Equation 2.

Based on these two sets of information, the peer p_i computes the utility value of each peer in LC_i using the equation 6. Depending upon its own its own node capacity, peer p_i selects a certain number of peers from the list LC_i and adds them into its neighbor list ($Nbr(p_i)$). The chances a peer $p_k \in LC_i$ being added to the neighbor list of p_i is directly proportional to p_ks utility values. Concretely, the probability of p_k being selected as a neighbor of p_i is given by the following equation.

$$P_i(LC_i, p_j) = r_i^{-\ln(r_i)} \cdot \frac{f_i(p_j) - r_i}{\sum_{p_k \in LC_i} f_i(p_k) - r_i} + (1 - r_i^{-\ln(r_i)}) \cdot \frac{\frac{1}{d_i(LC_i, p_j)} - (1 - r_i)}{\sum_{p_k \in LC_i} \frac{1}{d_i(LC_i, p_k)} - (1 - r_i)} \quad (6)$$

The peer p_i now sets up its outgoing edges (forwarding connections) to each node in its neighbor list. It then initiates the process to setup the incoming edges (back links to p_i) by sending a backward connection request to each peer $p_k \in Nbr(p_i)$. The request is augmented with the capacity information C_i of peer p_i and its network coordinates. A peer receiving a backward connection request utilizes a similar utility principle to decide whether to accept the request. This ensures that powerful peers are easily accepted by other powerful peers as their neighbors whereas weaker ones are good candidates only when they are close enough. The GroupCast system also includes an epoch-based scheme for maintaining the structure of the P2P overlay even when the network experiences significant churn [27].

4 Experimental Evaluation

We have implemented a discrete event simulation system to evaluate GroupCast. This system is an extended Java version of p-sim [19] system. We used the Transit-Stub graph model from the GT-ITM topology generator [26] to simulate the underlying IP networks. Peers are randomly attached to the stub domain routers and organized into overlay networks using the algorithm presented in Section 3.3. The capacity of peers is based on the distribution gathered in [23],

Table 1. Capacity distribution of peers

Capacity level	Percentage of peers
1x	20%
10x	45%
100x	30%
1000x	4.9%
10000x	0.1%

as shown in Table 1. We use GNP [1] to assign network coordinate to each peer. Each experiment is repeated over 10 IP network topologies. Each IP network supports 10 overlays, and each overlay network provides service for 10 communication groups.

4.1 Evaluating the GroupCast Service Lookup Mechanism

We begin by evaluating the utility-aware spanning tree construction and group communication mechanisms of the GroupCast system. Most unstructured use either scoped flooding (broadcast) or random walk as their communication paradigm. However, flooding-based mechanisms are expensive in terms of message loads they impose, whereas random walks result in longer delays. The GroupCast system includes a selective service announcement (SSA) mechanism for efficient and low-cost service lookups.

The first experiment evaluates the effectiveness and efficiency of the SSA scheme by simulating the service announcement processes in a number of overlay networks that are generated using either our utility-aware overlay construction mechanism or the centralized PLOD algorithm. In order to gain a better understanding, we compare the SSA mechanism with the non-selective service announcement (NSSA) scheme (see Section 2.1). For each overlay network, we randomly select 10 peers as rendezvous points, and initiate the selective service announcement (SSA) process and the non-selective service announcement (NSSA) process from each of them. For both SSA and NSSA, we first record the fraction of peers in the overlay that have received the service announcement. As we mentioned earlier, when these peers want to subscribe for the group communication service, they can circumvent the service searching process. For peers that have not received the service announcement message, subscription process involves searching its neighborhood for peers that have received the service announcement message. In our simulator, these peers use a ripple flooding search scheme for this purpose with TTL being set to 2. We measure the success rates of service lookups for both SSA and NSSA schemes. We also record the total number of messages generated by these two schemes.

The results in Figure 5 show that the SSA scheme reduces the total number of messages generated in both GroupCast and random power-law overlay networks. The SSA scheme limits the number of subscription messages sent to neighbors that are not likely to be a part of the communication group. This reduces the message load by 63% to 70% when compared with NSSA scheme for the GroupCast overlay. The reduction is 35% to 44% for the random power-law overlay. We notice that the number of subscription messages of SSA scheme in random power-law overlays is almost negligible. This is because GroupCast overlays have lower cluster coefficient values than the random power-law topologies generated using PLOD. Thus, SSA messages reach fewer peers. The results also show that the SSA scheme performs better for networks with higher connectivity value.

Figure 6 leads to two interesting observations. First, fewer peers in GroupCast receive the SSA messages compared to random power-law topology. Second, all subscribers can locate their services with 100% success rate even when the initial

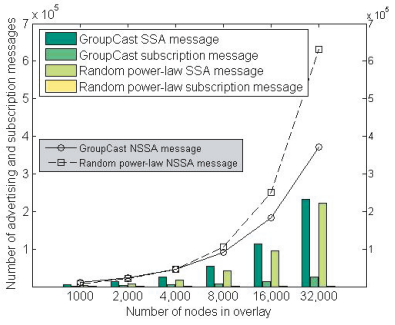

Fig. 5. Message loads of service lookup schemes

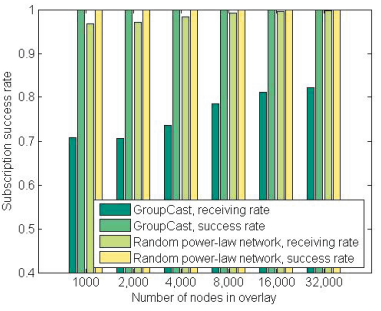

Fig. 6. Success rate of service lookup in GroupCast overlay and random power-law overlay with SSA

TTL of the subscription messages is set to *two*. This is essentially because, in the GroupCast overlay, the neighbors of individual peers are likely to have higher utility values. Hence, at each step of the SSA process, more candidate peers meet utility-aware selection criterion. This is also the reason for the relatively large number of service announcement messages in the GroupCast overlay when compared with random power-law network. However, the peers chosen by our utility-aware selection mechanisms are more suitable to the group communication spanning trees and they ensure high subscription success rates even at very small TTL values.

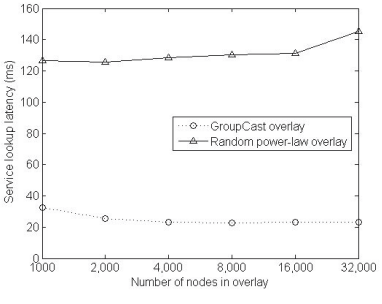

Fig. 7. Latency of service lookup in GroupCast overlay networks and random power-law overlay networks using selective service announcement

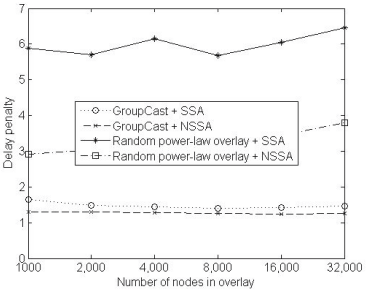

Fig. 8. Delay penalty of group communication applications

4.2 Improvement of Application Performance

The second set of experiments studies the effects of the proposed techniques on a group communication application. The group communication application we consider is that of end-system multicasting (ESM). ESM has been proposed as

an alternative for IP multicast, which has suffered from lack of wide acceptance and deployment. In this approach, peers form overlay networks and implement multicast functionality. Multicast data are replicated on peers and propagated through unicast edges of the overlay networks. ESM is inherently less efficient than IP multicast, as ESM may send packets with same payload multiple times over the same physical network link. Moreover, the ESM workload distribution among heterogeneous peers affects the overall system performance.

We simulated P2P overlay networks consisting of 1×10^3 to 3.2×10^4 peers. P2P overlay networks are constructed using our utility-aware mechanism as well as the centralized PLOD algorithm. We used the routing weights generated by the GT-ITM package to simulate the IP unicast routing. IP multicast systems are simulated by merging the unicast routes into shortest path trees. We use both SSA and NSSA for service announcement and subscription management.

We quantify the performance of the schemes using *Relative Delay Penalty* and *Link Stress* parameters, which are the two popular metrics for evaluating the efficiency of ESM systems. Relative delay penalty is defined as the ratio of the average ESM delay to the average IP multicast delay. Link stress is defined as the ratio of the number of IP messages generated by an ESM tree to the number of IP messages generated by an IP multicast tree interconnecting the same set of subscribers.

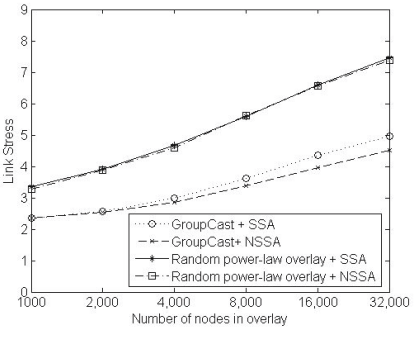

Fig. 9. Link stress of group communication applications

Figure 8 shows the relative delay penalties when multicasting is implemented through various combinations of the two overlay management schemes (utility-aware (GroupCast) and random power-law) and the two spanning tree construction schemes (SSA and NSSA). Figure 9 shows the respective link stress values. The results show that ESM implemented on GroupCast overlays yield significant improvements in terms of both metrics when compared with their counterparts implemented on random power-law networks. The delay penalty of ESM implemented on GroupCast overlay is around 1.5, which is close to the theoretical lower bound of 1. The link stresses of ESM implemented on GroupCast is about two-thirds the link stresses of ESM implemented on top of random power-law network. The improvements are due to the network proximity awareness of the

GroupCast overlay networks. Multicast payloads are forwarded through shorter paths, thus generating fewer IP packets in the underlying IP network.

It is interesting to note that the impact of the SSA scheme on application performance is almost negligible in GroupCast overlay networks, whereas the impact in random power-law networks is significant. We attribute this behavior to the fact that GroupCast overlay networks are already aware of the network proximity of peers. Thus, the peers chosen by the SSA scheme are most likely be the ones that are actually used in the information dissemination spanning tree.

5 Related Work

The work on group communication in P2P networks has mainly focused on structured P2P networks. Researchers have proposed several application-level multicasting schemes for DHT-based structured overlay networks [20,22,24]. However, structured P2P networks have high maintenance costs, especially in highly dynamic environments. In contrast, the GroupCast system does not require any DHT abstractions from the overlay. Instead, Our techniques are completely distributed, and they rely only on local information.

Many distributed group communication systems rely on the services of overlay networks for operation [7,11,14]. Usually, end-hosts in the communication groups use the unicast links of overlay networks to exchange application and management messages. Researchers have explored various techniques to optimize the system performance at the application level with the objective of designing efficient and scalable query processing mechanisms [13].

A popular approach to improving P2P networks is to utilize the rankings of different peers in terms of their node capacity and organize them into different hierarchical layers [4,25]. However, such predetermined hierarchical structures can introduce system vulnerabilities. Further, for efficiency purposes, the supernodes maintain the state information of the normal peers they serve. However, such state information is generally tied to the application, and it is hard to design a supernode overlay layer that can serve as a generic middleware to support different services.

Adaptation mechanisms have been studied in the context of application-layer multicasting [8,28]. Our research is complimentary to these works. These systems can utilize the GroupCast protocols for constructing well-regulated spanning trees. Our protocol can help reduce the number of adaptations by ensuring the efficiency of initial spanning trees. Techniques such as RON [6] have been designed for building generic overlays that are independent of the applications built on top of them. However, unlike these works, the GroupCast system constructs overlay networks that incorporate network proximity information, and it also builds *scale-free* power-law topologies assigning connections according to the peers' capacities.

In short, the work presented in this paper has several unique features and our system addresses a problem that is crucial for the success of several multi-party collaborative applications.

6 Conclusion

This paper presents the design and evaluation of *GroupCast* – a utility-aware decentralized middleware architecture for scalable and efficient wide-area group communications. The GroupCast design incorporates three novel features: (a) A *utility function* that measures the usefulness of unicast links to the scalability and efficiency of the group communication application; (b) A distributed utility-aware scheme for constructing efficient spanning trees for disseminating group communication payloads; and (c) A utility-based overlay management protocol for generating and maintaining low-diameter overlay networks. Our experiments show that GroupCast provides an order of magnitude improvement in the scalability of wide-area group communication applications.

Acknowledgement

This work is partially supported by grants from NSF CSR, NSF SGER, NSF CyberTrust, and grants from IBM SUR and IBM Faculty Award, DoE SciDAC, and AFOSR.

References

1. E. Ng. GNP software (July 2006), `http://www.cs.rice.edu/eugeneng/research/gnp/software.html`
2. Google Talk (July 2006), `www.google.com/talk/`
3. Live Meeting (July 2006), `http://www.microsoft.com/office/livemeeting`
4. Sharman networks LTD. KaZaA media desktop (July 2006), `http://www.kazaa.com`
5. Skype. (July 2006), `http://www.skype.com`
6. Andersen, D., Balakrishnan, H., Kaashoek, F., Morris, R.: Resillient overlay networks. In: SOSP. Proceedings of the 18th ACM Symposium on Operating Systems Principles, ACM Press, New York (2001)
7. Banerjee, S., Bhattacharjee, B., Kommareddy, C.: Scalable application layer multicast. In: Proceedings of the 2002 ACM SIGCOMM Conference, ACM Press, New York (2002)
8. Banerjee, S., Kommareddy, C., Kar, K., Bhattacharjee, B., Khuller, S.: Construction of an efficient overlay multicast infrastructure for real-time applications. In: Proceedings of INFOCOM (2003)
9. Baset, S.A., Schulzrinne, H.: An analysis of the skype peer-to-peer internet telephony protocol. Technical Report cucs-039-04, Dept. of Computer Sci., Columbia Univ. (2004)
10. Bu, T., Towsley, D.: On distinguishing between internet power law topology generators. In: IEEE INFOCOM, New York, NY, IEEE, Los Alamitos (2002)
11. Castro, M., Druschel, P., Kermarrec, A., Rowstron, A.: SCRIBE: A large-scale and decentralized application-level multicast infrastructure. IEEE Journal on Selected Areas in communications (JSAC) (2002)
12. Chawathe, Y.: Scattercast: An Architecture for Internet Broadcast Distribution as an Infrastructure Service. PhD thesis, University of California, Berkeley (2000)

13. Chawathe, Y., Ratnasamy, S., Breslau, L., Lanham, N., Shenker, S.: Making Gnutella-like P2P systems scalable. In: ACM SIGCOMM, Karlsruhe, Germany, ACM Press, New York (2003)
14. Chu, Y.-H., Rao, S.G., Zhang, H.: A case for end system multicast. In: ACM SIGMETRICS 2000, pp. 1–12. ACM Press, New York (2000)
15. Dabek, F., Cox, R., Kaashoek, F., Morris, R.: Vivaldi: A decentralized network coordinate system. In: ACM SIGCOMM, Portland, Oregon, USA, ACM Press, New York (2004)
16. Deering, S., Cheriton, D.: Multicast routing in datagram internetworks and extended lans. ACM Transactions on Computer Systems 8(2) (May 1990)
17. Francis, P.: Yoid: Extending the multicast internet architecture (1999)
18. Jannotti, J., Gifford, D.K., Johnson, K.L., Kaashoek, M.F., O'Toole Jr., J.W.: Overcast: Reliable multicasting with an overlay network. In: OSDI. Proceedings of Symposium on Operating System Design and Implementation, pp. 197–212 (2000)
19. Merugu, S., Srinivasan, S., Zegura, E.: p-sim: A simulator for peer-to-peer networks. In: MASCOTS 2003. Proc. of the 11th IEEE Intl. Symp. on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems, IEEE Computer Society Press, Los Alamitos (2004)
20. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A scalable content addressable network. In: Proceedings of SIGCOMM, ACM Press, New York (2001)
21. Ratnasamy, S., Handley, M., Karp, R., Shenker, S.: Application-level multicast using content-addressable networks. In: Crowcroft, J., Hofmann, M. (eds.) NGC 2001. LNCS, vol. 2233, Springer, Heidelberg (2001)
22. Rowstron, A., Druschel, P.: Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In: Guerraoui, R. (ed.) Middleware 2001. LNCS, vol. 2218, pp. 329–350. Springer, Heidelberg (2001)
23. Saroiu, S., Gummadi, P., Gribble, S.: A measurement study of Peer-to-Peer file sharing systems. In: Proceedings of MMCN, San Jose, CA (August 2002)
24. Stoica, I., Morris, R., Karger, D., Kaashoek, F., Balakrishnan, H.: Chord: A scalable Peer-To-Peer lookup service for internet applications. In: Proceedings SIGCOMM, pp. 149–160. ACM, New York (2001)
25. WWW. The Gnutella RFC (July 2006), `http://rfc-gnutella.sourceforge.net`
26. Zegura, E.W., Calvert, K.L., Bhattacharjee, S.: How to model an internetwork. In: IEEE Infocom, vol. 2, pp. 594–602. IEEE, Los Alamitos (1996)
27. Zhang, J., Liu, L., Ramaswamy, L., Pu, C.: Scalable and Efficient Peer-to-Peer Group Communication: A Utility-based Approach. Technical report, College of Computing, Georgia Tech. (2006)
28. Zhou, Y., et al.: Adaptive reorganization of conherency-preserving dissemination tree for streaming data. In: ICDE (2006)