Characterizing Overlay Multicast Networks

Sonia Fahmy and Minseok Kwon
Department of Computer Sciences, Purdue University
e-mail: {fahmy,kwonm}@cs.purdue.edu *

Abstract

Overlay networks among cooperating hosts have recently emerged as a viable solution to several challenging problems, including multicasting, routing, content distribution, and peer-to-peer services. Application-level overlays, however, incur a performance penalty over router-level solutions. This paper characterizes this performance penalty for overlay multicast trees via experimental data, simulations, and theoretical models. Experimental data and simulations illustrate that (i) the average delay and the number of hops between parent and child hosts in overlay trees generally decrease, and (ii) the degree of hosts generally decreases, as the level of the host in the overlay tree increases. Overlay multicast routing strategies, together with *power-law and small-world Internet topology characteristics, are causes of the observed phenomena. We compare three overlay multicast protocols with respect to latency, bandwidth, router degrees, and host degrees. We also quantify the overlay tree cost. Results reveal that* $\frac{L(n)}{U(n)} \propto n^{0.9}$ *for small* n*, where* $L(n)$ *is the total number of hops in all overlay links,* $U(n)$ *is the average number of hops on the source to receiver unicast paths, and* n *is the number of members in the overlay multicast session.*

1. Introduction

Overlay networks have recently gained attention as mechanisms to overcome deployment barriers to router-level solutions of several networking problems. Overlay solutions for multicasting [7, 13, 15, 21, 16], inter-domain routing pathologies [3], content distribution and content sharing [22] are being extensively studied. In this paper, we consider a number of *overlay* (application-layer) multicast approaches which have been proposed over the last three years. In overlay multicast, hosts participating in a multicast session form an overlay network, and only utilize unicasts among pairs of hosts (considered neighbors in the overlay tree) for data dissemination. The hosts in overlay multicast exclusively handle group management, routing, and tree construction, without any support from Internet routers.

The key advantages overlays offer are flexibility, adaptivity, and ease of deployment. Overlays, however, impose a performance penalty over router-level alternatives. While overlay multicast clearly consumes additional network bandwidth and increases latency over IP multicast, little attention has been paid to precisely quantifying this overlay performance penalty, either theoretically or experimentally. Moreover, to the best of our knowledge, there is no work on characterizing overlay multicast tree structure. Such characterization is important to gain insight into overlay properties and their causes at *both* the application layer and the underlying network layer. It is also important to compare different overlay multicast strategies to determine how to meet the goals of target applications (e.g., by balancing latency versus bandwidth tradeoffs).

In this paper, we analyze overlay multicast trees via (i) real data integrated from End System Multicast (ESM)/Narada [7] experiments and traceroute servers, (ii) simulations of three representative classes of overlay multicast strategies, and (iii) simple analytical models. We quantify several aspects of the performance penalty associated with overlay multicast, with emphasis on the overlay cost (i.e., efficiency) at the network-layer. We derive and validate asymptotic forms of the overlay cost from two different tree models.

Our results indicate that (i) the average delay and the number of hops between parent and child hosts generally decrease, and (ii) the degree of hosts generally decreases, as the level of the host in the overlay tree increases. We find that overlay multicast routing strategies, *together with* power-law and small-world Internet topology characteristics, are causes of these observed phenomena. We isolate the impact of each of these causes. Our results also reveal that $\frac{L(n)}{U(n)} \propto n^{0.9}$ for small n, where $L(n)$ is the total number of hops in all overlay links (connections), $U(n)$ is the average number of hops on the source to receiver unicast

* This research has been sponsored in part by the Purdue Research Foundation, and the Schlumberger Foundation technical merit award.

paths, and n is the number of members in the overlay multicast session. This can be compared to an IP multicast cost proportional to $n^{0.6}$ to $n^{0.8}$ [6, 8].

The remainder of this paper is organized as follows. In Section 2, we describe overlay networks and their performance metrics. In Section 3, we characterize overlay multicast networks via simulations and experimental data analysis. In Section 4, we propose and validate an overlay multicast model based on our observations. In Section 5, we discuss related work. Finally, we summarize our conclusions and future work in Section 6.

2. Overlay Networks: Definitions and Metrics

We consider the *underlying network* as a graph $G = (N, E)$, where N is a set of nodes, and E is a set of edges. A node $\eta_i \in N$ denotes a *router*, and an edge $(\eta_i, \eta_j) \in E$ denotes a bi-directional physical link in the underlying network. An *overlay network* superimposed on G is a *tree* $o = (s, D, N_o, E_o)$, where s is the source host, D is the set of receiver hosts, $N_o \subseteq N$ is the set of nodes in the underlying network G that are traversed by overlay links, and E_o is the set of overlay links, defined below.

The set of hosts H_o consists of s and D in o, i.e., $H_o = \{s\} \cup D$. The cardinality of set H_o is equal to n. An overlay link $e_o = (d_s, \eta_0, \ldots, \eta_{ls}, d_r) \in E_o$ comprises a host $d_s \in H_o$, followed by a sequence of routers $\eta_i \in N_o$, followed by a host $d_r \in D$. Each receiver $\in D$ appears exactly once at the *end* of any sequence denoting an overlay link, but may appear multiple times at the *beginning* of sequences for different overlay links. An overlay link is typically a UDP or TCP connection established by the overlay multicast protocol.

The number of hops in the router sequence $\eta_0, \ldots, \eta_{ls}$ in an overlay link $e_o \in E_o$ is denoted by ls. For every two routers $\eta_i, \eta_j \in N_o$ that appear consecutively in an overlay link $e_o \in E_o$, there must exist a link connecting them in the underlying network, i.e., edge $(\eta_i, \eta_j) \in E$ holds. The same router $\eta_i \in N_o$ can appear in multiple overlay links $e_o \in E_o$. Subsequences of routers $\eta_i, \ldots, \eta_j$ can also appear in multiple overlay links $e_o \in E_o$. Figure 1 illustrates an example overlay network with 6 overlay links.

Given an overlay network o, we define the term *overlay cost* as the number of underlying hops traversed by every overlay link $e_o \in E_o$ for an overlay o. More formally, the overlay cost is: $\forall e_o \in E_o, \Sigma \ ls(e_o)$, where $ls(e_o)$ denotes the number of router-to-router hops between $\eta_0, \ldots, \eta_{ls}$ for the overlay link e_o (as defined above). We consider the first and last hops to/from hosts separately. This is because we must fairly compare the normalized overlay cost to the normalized IP multicast cost computed in [8, 18, 2], where the first and last hops are ignored. For example, the overlay cost for the overlay in Figure 1 is 2+3+1+1+4+2=13.

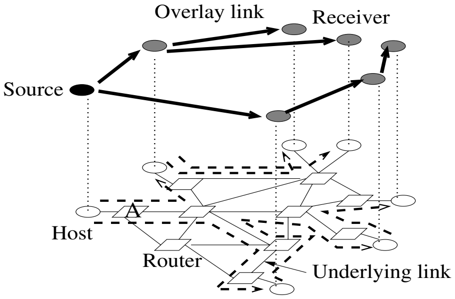


Figure 1. An example overlay multicast tree over an underlying network

We also use the term *link stress* to denote the total number of identical copies of a packet over the same underlying link (as defined in [7]). For example, the stress of the link from the source to A in Figure 1 is two. It is clear that the overlay cost defined above can be represented as $\forall i, \sum_i stress(i)$ where i is any *router-to-router* link traversed by one or more overlay links $e_o \in E_o$, and $stress(i)$ is the stress of link i. Prior work also used a "resource usage" metric, defined as $\forall i, \sum_i delay(i) \times stress(i)$, where i is an underlying link traversed by one or more overlay links [7]. Our overlay cost metric is a special case of this resource usage notion, when $delay(i) = 1, \forall i$. We opt to evaluate delays separately from the overlay cost, in order to isolate the delay and stress aspects of an overlay.

In addition to the overlay cost and link stress, we study the following overlay tree metrics: (1) degree of hosts H_o (equivalent to the host contribution to link stress of the host-to-first-router link), (2) degree of routers $\in N_o$, and hop-by-hop delays of underlying links traversed by overlay links $\in E_o$, (3) overlay tree height, (4) delays and number of hops between parent and child hosts, (5) mean bottleneck bandwidth between the source s and receivers $\in D$, and (6) mean latency, longest latency, and relative delay penalty (RDP) from the source to a receiver.

The latency $latency(s, d_r)$ from the source s to $d_r \in D$ is: $delay(s, d_0) + \sum_{i=0}^{l-1} delay(d_i, d_{i+1}) + delay(d_l, d_r)$, assuming s delivers data to d_r via the sequence of hosts $(d_0, \cdots, d_l)$. Here, $delay(d_i, d_{i+1})$ denotes the end-to-end delay of the overlay link from d_i to d_{i+1}, for $d_i \in H_o$ and $d_{i+1} \in D$. Note that the RDP from s to d_r (defined in [7]) is the ratio $\frac{latency(s,d_r)}{delay(s,d_r)}$. We compute the mean RDP of all receivers $\in D$. We can also define the *stretch* as $\frac{hops(s,d_r)}{ls(s,d_r)+2}$ where $hops(s, d_r) = ls(s, d_0) + \sum_{i=0}^{l-1}(ls(d_i, d_{i+1}) + 2) + ls(d_l, d_r) + 4$. Stretch denotes the relative number of hops instead of the relative latency used in RDP. These metrics compare overlay multicast to unicast (or IP multicast using a minimum delay tree). It is clear that there is a tradeoff between the latency metrics and the stress/bandwidth metrics. Balancing this tradeoff is the key to effective overlay multicast protocol design.

3. Overlay Multicast Tree Structure

Our primary goal in this section is to isolate the impacts of (i) the overlay protocol, (ii) the underlying network connectivity and routing, and (iii) the overlay host distribution, on the overlay tree structure. We first analyze experimental data, and then conduct a set of simulations.

3.1. Experimental Data

In order to study the structure of *real* overlay networks in the Internet, we analyze recent experimental results for the End System Multicast (ESM) protocol [7]. We recorded the overlay trees constructed during experiments performed by the ESM developers in November 2002. (Unfortunately, the ESM developers have not released the overlay tree structure in their later experiments.) Since the overlay trees did not change significantly throughout the experiment lifetime, we selected one representative overlay tree. The tree comprises 65 hosts. We are currently setting up ESM on the PlanetLab testbed [17] to collect more measurements.

We use *traceroute* to find the underlying path between every two hosts on the overlay tree. We encountered two problems using traceroute. First, some routers do not generate ICMP Time-Exceeded packets when TTL (Time-To-Live) reaches zero. Second, many routers disable the source-route capability, primarily due to security concerns. Due to this, finding paths between two arbitrary hosts via traceroute (without having accounts on either of these hosts) becomes difficult. We utilize publicly available traceroute servers [1] and our own machines to compute paths to all the hosts on the overlay tree. These paths are then synthesized to approximate the paths between any two overlay hosts. Our task was simplified because the hosts in the experiments, with a few exceptions, are located at universities in the United States. Most university hosts are connected to the Internet2 backbone network [12], and thus the routes typically intersect at points on Internet2. These points provide the synthesis junctions used for path extraction.

Figure 2 depicts the mean number of hops between every two parent-child ESM hosts, for hosts at different levels of the overlay tree (90% confidence intervals are shown to indicate variability). The figure shows that the number of hops typically decreases as the host level increases, though the decrease is not monotone. We now seek the causes of this phenomenon. Consider a set of routers that are connected according to the power-law [10] and small-world [4] properties. The power-law property dictates that there is a larger number of low-degree routers than high-degree routers. We surmise that a high-degree high-bandwidth router is typically more likely to be traversed by overlay links near the source of the overlay tree. This is because a high-degree router has higher chances of reducing the path length and delays than a low-degree router, due to its connectivity to a larger number of routers. The high-degree router is also more likely to have high bandwidth links connected to it. Overlay multicast protocols which consider delay, path length, or bandwidth are thus likely to exploit such high-degree routers in the first few levels of the tree (unless all hosts are clustered near the source). Recall also that nearby hosts tend to be clustered by the small-world property. Accordingly, we can visualize an overlay tree where a number of high-degree routers connect the hosts at the first few levels of the tree. In addition, many hosts are connected to low-degree lower-bandwidth routers, which are clustered at lower levels of the tree. Therefore, hosts at lower levels of the overlay tree may only be a few hops away from each other.

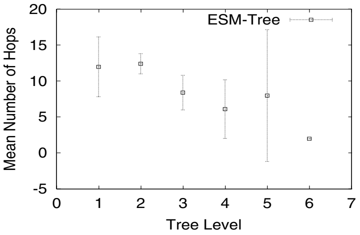
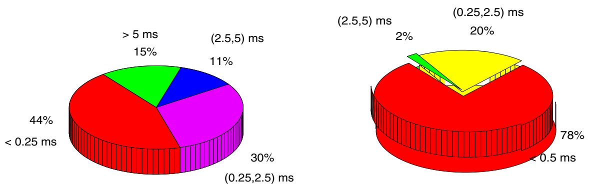


Figure 2. Number of router-to-router hops between parent-child ESM hosts versus level of host in overlay tree

Figure 3 shows the distribution of per-hop delay (the delay between two consecutive routers on a path from a parent to a child ESM host) for different overlay tree levels. The per-hop delay between two consecutive routers η_i and η_j is estimated as $\frac{1}{2}rtt(\eta_i, \eta_j)$, where $rtt(\eta_i, \eta_j)$ is the time to travel from η_i to η_j and vice versa obtained via traceroute. The figure indicates that 78% of per-hop delays in lower tree levels (levels 4-6) are shorter than 0.25 ms, and only 2% are between 2.5 and 5 ms. In contrast, only 44% of per-hop delays are shorter than 0.25 ms, and 15% exceed 5 ms, for the first level of the tree, which agrees with our earlier explanation. The round trip times between every two parent-child ESM hosts at different levels of the overlay tree were also found to generally decrease as the host level increases, confirming our intuition. Finally, we have found that the degree of hosts in the overlay tree grows as hosts get closer to the root of the overlay tree. This decreasing degree can be attributed ESM's goal of minimizing delay (if bandwidth is acceptable).

3.2. Simulation Experiments

We also investigate the overlay structure via simple session-level simulations.



(a) Tree level 1

(b) Tree level 4-6

Figure 3. Distributions of per-hop delay for different ESM overlay tree levels

3.2.1. Simulation Setup Our simulation topology contains 4000 routers connected according to power-law and small-world properties. In a power-law graph, a complementary cumulative distribution function $cd^{-\alpha}$ is used to denote the fraction of routers with degree greater than d, where c and α are constants [14]. We use $c = 1$ and $\alpha = 1.22$. Groups of routers are clustered according to the small-world property: a router connects to its closest neighbor routers with probability p, and to other routers with probability $1 - p$, according to router degree. We use $p = 0.5$. Routers are uniformly distributed on a 500×500 plane, and the Euclidean distance between two routers approximates the delay between the two routers (in ms). Hosts are connected to edge routers (which are defined as routers with degree less than 10) uniformly at random. The bandwidth from edge routers to hosts is selected according to the realistic distribution: 40% are 56 kbps, and 15% for each of 1.5, 5, 10, 100 Mbps. All other links are assigned bandwidths ranging from 100 Mbps to 1 Gbps. The underlying network routes are selected to optimize *delays*. It is also worth mentioning that we have simulated smaller scale topologies and the results were similar. Results for Transit-Stub topologies generated by the popular GT-ITM can be found in [9].

We simulate three representative overlay multicast protocols on the two topologies: ESM [7], Topology-Aware Grouping (TAG) [15], and Minimum Diameter Degree-Bounded Spanning Tree (MDDBST) [21]. The reason we select ESM is that it is the first overlay multicast protocol to be widely tested in the Internet. Each ESM host evaluates the utility of other hosts to determine its neighbors. A host has an upper degree bound (UDB) on the number of its neighbors. We use a value of 6 for the upper degree bound. The ESM flavor used in our simulations has two discretized bandwidth levels: $>$ 100 kbps and $\leq$ 100 kbps (same as the version used for multicasting SIGCOMM 2002). The overlay tree is first optimized for bandwidth, and then uses delay as a tie breaker among hosts at the same bandwidth level.

The second class of protocols we investigate is topology-aware overlay multicast protocols, which includes Scribe [5], topology-aware Content-Addressable Network (CAN) [20], and TAG [15]. We select TAG as a representative of this group. TAG is a simple and faithful representation of topology-based approaches, since it aligns overlay routes and underlying routes, if bandwidth constraints are met. A TAG host becomes the child of the host that most "matches" its path. Here, a path is defined as the sequence of routers from the source to the host. A's path matches B's path when the path from the source to A is a prefix of the path from the source to B. This flavor of TAG is called "complete path matching." We use the partial path matching version, which allows u unmatched routers at the end of the prefix. Partial path matching is activated when the bandwidth from a potential parent to a new member is less than a threshold $bwthresh$. We use $u = 3$ and $bwthresh = 20$ kbps in our simulations.

The third class of protocols we investigate includes protocols that seek to minimize overlay cost [16], or the longest path in an overlay network [21] (with delay or bandwidth constraints). We select MDDBST, given in [21], as a representative protocol in this class. MDDBST minimizes the number of hops in the longest path, and bounds the degree of hosts. We use a degree bound of 20 in our simulations.For each protocol, we run five simulations with different random number generator seeds (for topology generation and for selecting the multicast source and destinations) and average the results.

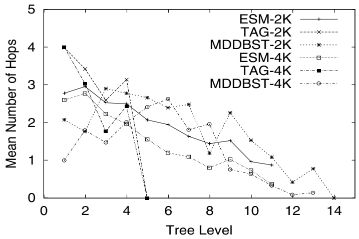


Figure 4. Number of hops versus overlay tree level in simulations on the power-law and small-world topology

3.2.2. Simulation Results Figure 4 illustrates the mean number of hops between parent and child hosts for different host levels in the overlay tree. The labels "ESM-2K" and "ESM-4K" denote ESM with 2000 or with 4000 members respectively, and so on. The figure reveals that the number of hops between parent and child hosts tends to decrease as the level in the overlay tree increases, for both ESM and TAG. MDDBST does not exhibit a clear trend. This is because MDDBST does not seek the shortest path to individual hosts, but minimizes the longest path. In general, the decreases are more pronounced for TAG than for the other two protocols, since TAG aligns overlay and underlying routes, subject to bandwidth availability. The observed decrease in mean number of hops is consistent with our experimental data, and our intuition about the effects of Internet topology characteristics.

In order to isolate the effects of the power-law property from the small-world property, we have run the same simulations on only-power-law (but no clustering) and only-small-world (but equal degree routers) topologies. The results revealed that *both* clustering among closely located routers as dictated by the small-world property, and power-laws of router degrees, contribute to the observed decrease in number of hops with overlay tree level increase. However, the effect of the power-law property is more dominant.

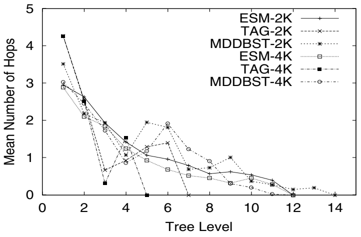


Figure 5. Number of hops versus overlay tree level in simulations on the power-law and small-world topology with non-uniform host distribution

We also simulate the three protocols with a *non-uniform* host distribution. In this case, we randomly select an edge router and then connect ω hosts to this router and its neighboring routers (one host per router), where ω is a random number between 1 and 20. Figure 5 illustrates that the number of hops between parent and child hosts decreases even more rapidly (though with some fluctuations) than uniform host distribution case (Figure 4). The decrease was less pronounced when we repeated the same experiment on the GT-ITM topology. Therefore, the power-law and small-world properties, and the non-uniform host distribution are all factors that exacerbate this phenomenon. The routing features of overlay multicast protocols, such as the utility for selecting neighbors in ESM, or topology awareness in TAG, also play an important role.

To validate our argument that high-degree routers tend to be traversed in upper levels of the overlay tree, we have also studied the average router degree versus the overlay tree level. As expected, higher degree routers are traversed at upper overlay tree levels. We also investigated the frequency that routers with certain degrees are traversed by overlay links. We found that all three protocol trees cross a significant number of high-degree routers (50+), in order to exploit their high connectivity and high bandwidth.

In addition, we have investigated the host degree versus the host overlay tree level. The host degree remains within a small range (≤ 20), except for the source host for the TAG protocol. This is because TAG attempts to send more copies from the source to reduce delay when all receivers are far from each other. As a result, the ESM and MDDBST trees are typically longer than TAG trees. The tree height increases as the number of members is increased, but the increase is slow beyond a certain number of members. We have also studied the total stress for all three protocols, and found that ESM exhibits the lowest stress, followed by MDDBST, then TAG.

Figure 6(a) depicts the relative delay penalty (RDP) (defined in Section 2) for the three protocols. ESM achieves the lowest RDP, except when the number of members is small. ESM, however, exhibits the highest longest latency (Figure 6(b)). The latencies and RDP for ESM decrease as more hosts join (especially since they are randomly located), because lower latency paths become available. In contrast, TAG RDP is high because partial path matching with bandwidth constraints increases latency when a large number of members join.

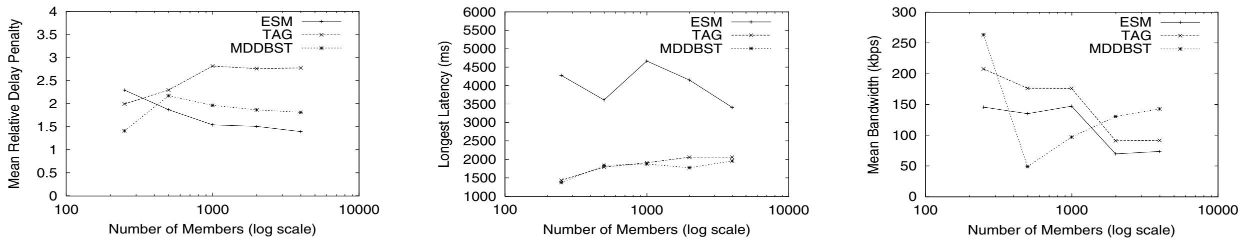
Finally, the mean bottleneck bandwidth between the source and receivers for all three protocols is illustrated in Figure 6(c). The receivers in TAG obtain a higher bandwidth than the receivers in ESM because of the TAG partial path matching. The bandwidth of MDDBST fluctuates since the degree bound of 20 does not depend on bandwidth, even though it should. Figures 6(a) and 6(c) together illustrate the latency versus bandwidth tradeoff in overlay multicast protocols. Note that these results vary with protocol parameters. For example, TAG gives lower RDPs and lower bandwidths with a smaller u or a smaller $bwthresh$. MDDBST can also increase bandwidth with a lower degree bound, at the expense of longer latencies and RDPs.

4. Overlay Multicast Tree Cost

In this section, we model overlay multicast trees based on the overlay tree structure we have observed, and compute their costs.

4.1. Network Model

We model the underlying network as a graph $G = (N, E)$ and the overlay tree o as the tuple (s, D, N_o, E_o), as defined in Section 2. To simplify our analysis, we transform G into a complete k-ary tree $G' = (N, E, r)$ on which o is constructed, where N and E are the same as in G, and $r \in N$ is designated as the root router. s is the only host connected to r. Other hosts are connected to routers with equal probability in both G and G' to obtain D. The height of G' is h. To transform G into G', any cycle in G is broken by eliminating the edge on the cycle which no overlay link in o traverses. Such an edge typically exists when the overlay cost is minimized, which is the overlay we consider here, as given in Definition 1 below. In addition, we move the children of nodes whose degree is larger than k, along with the subtrees rooted at these nodes, to nodes which have degree less than k. Such nodes are guaranteed to exist, e.g.,



(a) Mean relative delay penalty

(b) Longest latency

(c) Mean bottleneck bandwidth

Figure 6. Mean relative delay penalty (RDP), longest latency, and mean bottleneck bandwidth tradeoffs in simulations on the power-law and small-world topology

leaf nodes. This simple transformation shows that we do not significantly lose generality by considering an underlying tree. The overlay cost exhibited with an underlying tree has also been shown to be more consistent with that exhibited with real topologies, compared to meshes or random graphs [19]. We are, however, currently investigating the average costs for the set of trees covering a power-law and small-world underlying network.

To incorporate the number-of-hops distribution properties discussed in Section 3, routers with only one child (and no hosts to be connected) are added between branching points in the underlying network model. Such routers are called *unary nodes*. We had observed that the number of hops between parent and child hosts approximately decreases, as the level of the host in the overlay tree increases. A similar modeling assumption to that in [2] (a *self-similar tree*) can be used to represent this observation. This entails that $A_i = \phi A_{i-1}$, $0 \leq \phi \leq 1$, where A_i is the number of concatenated links generated by unary nodes between a node at level $i-1$ and a node at level i in the underlying network (the notions of levels and h do not consider unary nodes, which are counted separately). Therefore, $k^{(h-i)\theta}-1$ unary nodes are created between adjacent nodes at levels $i-1$ and i. This implies that $k^{(h-i)\theta}$ links exist at level i from a branching node at level $i-1$. The tree has no unary nodes when $\theta = 0$. Note that the number of hops on overlay links will not be monotonically decreasing (but will be approximately decreasing) for increasing levels of the overlay tree, since data may be disseminated up G' in certain segments, as discussed in the next 2 sections.

We assume that each receiver is connected to a router in the network uniformly and independently of other receivers. We use the term $L_o(h,k,n)$ to denote overlay cost for an overlay tree o and number of hosts $|H_o| = n$ (h and k are defined above). In [8], m, the number of distinct routers to which hosts are connected, is used instead of n in $L_o(h,k,n)$. We, however, believe that using the number of hosts n is intuitively appealing and makes analysis simpler. Note that m can be approximated by $M(1-(1-\frac{1}{M})^n)$, where M is the total number of available routers to which hosts can be connected. Therefore, $m \approx n$ when $\frac{n}{M} \ll 1$ [2].

Among all possible overlay networks that can be superimposed on G', we compute the *least cost* overlay network defined as follows.

Definition 1 *Let Ω be the set of all possible overlays, connecting a particular set of n hosts, and superimposed on a network G'. Let $L_\tau(h,k,n)$ be the overlay cost for $\tau \in \Omega$. Let o be the least cost overlay on G'. Then, o is the overlay that satisfies $L_o(h,k,n) \leq L_\tau(h,k,n)$ for all $\tau \in \Omega$.*

We consider the least cost overlay network for three primary reasons. First, modeling and analysis are simplified in this case. Second, many overlay multicast protocols optimize a delay-related metric, which is typically also optimized by underlying (especially intra-domain) routing protocols. Third, it gives a lower bound on the overlay tree cost under our assumptions.

4.2. Receivers at Leaf Nodes

We first consider a network in which receivers can only be connected to leaf nodes in the underlying network. Figure 7(a) shows a model of such a network. One host, which is the current source of the overlay multicast session, is connected to the root r of the tree. All other hosts are connected to leaf nodes, selected independently and uniformly. We define ρ to be the lowest level with branching nodes above or at *half* of the tree height. Since $\sum_{i=\rho+1}^{h} k^{(h-i)\theta}$ indicates the height from ρ to the lowest tree level, ρ can be computed as: $2\sum_{i=\rho+1}^{h} k^{(h-i)\theta} \leq \sum_{i=1}^{h} k^{(h-i)\theta}$. Thus, $\rho = \left\lceil h - \frac{1}{\theta}\log_k \frac{k^{h\theta}+1}{2} \right\rceil$.

For ease of counting, we first consider a tree without unary nodes and then add the cost introduced by unary nodes. Figure 7(a) shows that the cost incurred when communicating from a receiver to another receiver, both connected to descendants of node σ at level $\lceil \frac{h}{2} \rceil$, is bounded by h. Otherwise, the source would send another copy directly to the receiver at cost h. For this reason, we group together all receivers connected to descendants of σ in a sub-

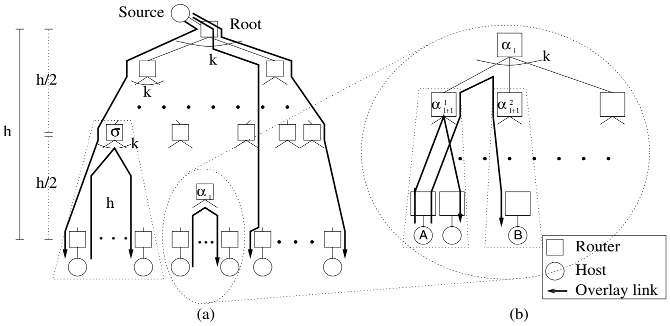


Figure 7. An overlay tree model with receivers located only at leaf nodes (for simplicity, unary nodes are not shown)

tree rooted at σ. Similar subtrees are created for every node at level $\lceil \frac{h}{2} \rceil$.

We divide the computation of $L_o(h,k,n)$ into two terms. The first term is the minimum cost to send to the subtrees rooted at σ, and the second term is the minimum cost of data dissemination within the subtrees. To compute the first term, we observe that there are k^ρ nodes at level ρ in the tree. The probability that a link connecting to level ρ is traversed by overlay o is $1-(1-k^{-\rho})^n$. Thus, the cost at level ρ is $k^\rho(1-(1-k^{-\rho})^n)$. Since $k^{(h-i)\theta}$ additional cost is incurred by a node at level i if the tree is extended with *unary nodes*, the first term becomes:

$$\sum_{i=1}^{h} k^{(h-i)\theta} k^\rho (1-(1-k^{-\rho})^n) = \frac{k^{h\theta}-1}{k^\theta-1} k^\rho (1-(1-k^{-\rho})^n) \qquad (1)$$

To compute the second term, we consider a subtree rooted at σ. This subtree and potential overlay links are shown in Figure 7(b). Consider a node α_l at level l, where $\frac{h}{2} \leq l < h$ in the subtree. Let α^0_{l+1} and α^1_{l+1} be two children of α_l at level $l+1$. Suppose that A is a receiver connected to a descendant of α^0_{l+1}, and B is a receiver connected to a descendant of α^1_{l+1}. Since $\sum_{i=l+1}^{h} k^{(h-i)\theta} \approx k^{(h-l-1)\theta}$ is incurred due to *unary nodes*, sending data from A to B across (up and then down) α_l costs: $2k^{(h-l-1)\theta}$.

Since there are k^{l+1} links to level $l+1$ of the tree, the probability that a particular link to level $l+1$ is used in o is $1-(1-k^{-(l+1)})^n$. Since α_l has k children, the cost from α_l to its children in o becomes $k(1-(1-k^{-(l+1)})^n)$. An overlay link is created between a pair of children of α_l, so the cost across α_l is $k(1-(1-k^{-(l+1)})^n)-1$. Therefore, for α_l, the cost at level l in the subtree becomes $2k^{(h-l-1)\theta}(k(1-(1-k^{-(l+1)})^n)-1)$. We, however, note that there must be no link across α_l if the cost from α_l to its children is less than one, that is, $k(1-(1-k^{-(l+1)})^n) < 1 \Leftrightarrow l > \ln_k(1-(1-\frac{1}{k})^{\frac{1}{n}})^{-1}-1$. Therefore, the cost at level l in the subtree $g(l)$ is defined as: $g(l) = 2k^{(h-l-1)\theta}(k(1-(1-k^{-(l+1)})^n)-1)$, where $\rho \leq l \leq \ln_k(1-(1-\frac{1}{k})^{\frac{1}{n}})^{-1}-1$. $g(l)=0$ otherwise. Consequently, the second term becomes: $\sum_{l=\rho}^{h-1} k^l g(l)$.

$L_o(h,k,n)$ is the sum of the first and second terms:

$$L_o(h,k,n) = \frac{k^{h\theta}-1}{k^\theta-1} k^\rho (1-(1-k^{-\rho})^n) + \sum_{l=\rho}^{h-1} k^l g(l) \quad (2)$$

We prove that this tree is indeed the least cost overlay tree on this underlying network in [9]. Since the average number of hops on the source to receiver unicast paths $U^\theta_o(h)$ is $\sum_{i=1}^{h} k^{(h-i)\theta} = \frac{k^{h\theta}-1}{k^\theta-1}$, the normalized overlay cost becomes: $R^\theta_o(h,k,n) = \frac{L_o(h,k,n)}{U^\theta_o(h)}$.

A power-law is observed in the normalized cost, where the exponent of n is $1-\theta$ (see [9] for details). Figure 8(a) depicts the normalized overlay cost $R^\theta_o(h,k,n)$ against the number of overlay group members n. The figure shows that $R^\theta_o(h,k,n) \propto n^{0.92}$, for $0 < a < 1$. Saturation occurs as $a \to \infty$ $(n \to \infty)$.

4.3. Receivers at Leaf or Non-leaf Nodes

We now relax the restriction that receivers are only connected to leaf nodes in the underlying network, as illustrated in Figure 9. A non-leaf node with receiver(s) connected receives data from an ancestor, and relays this data to its descendants. In contrast, descendants of a non-leaf node which has no receivers connected must receive data from other non-ancestor nodes.

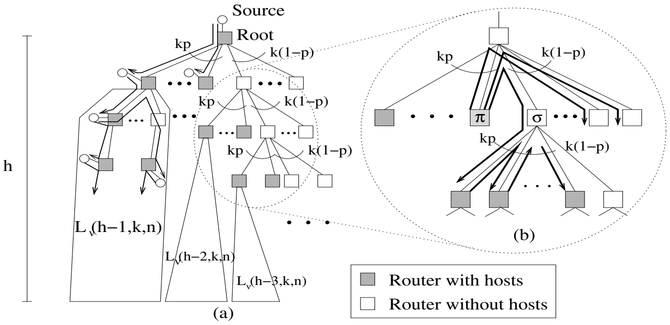
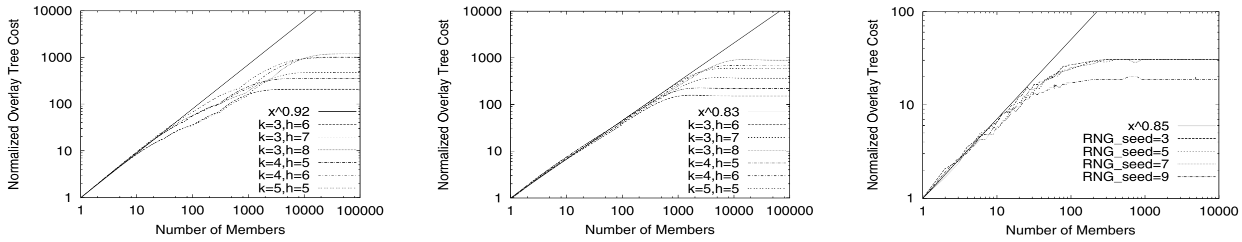


Figure 9. An overlay tree model with receivers located at leaf or non-leaf nodes (for simplicity, unary nodes are not shown)

We use the same underlying network model as in Section 4.2. We now assume that receivers are uniformly and independently distributed over the entire tree (with the exception of unary nodes). This implies that the probability that a node (other than the root) has at least one receiver connected is: $p = 1-(1-\frac{1}{M})^n$ for n receivers, where $M = k + \cdots + k^h = \frac{k^{h+1}-k}{k-1}$.

On the average, among the k children of a non-leaf node, kp children have receivers connected, while $k(1-p)$ children have no receivers connected. Let $L_\nu(h,k,n)$ be the



(a) Receivers at leaf nodes

(b) Receivers at leaf or non-leaf nodes

(c) Traceroute-based overlay multicast simulations using minimum spanning trees

Figure 8. Normalized overlay cost versus number of members from $R_o(h,k,n)$ for (a) and from $R_\nu(h,k,n)$ for (b) ($\theta = 0.1$) and from simulations for (c) (log-log scale)

overlay cost of an overlay network ν. The computation of $L_\nu(h,k,n)$ is split into two components: (i) cost for kp children of the root with receivers, and (ii) cost for $k(1-p)$ children of the root without receivers. Again, we first consider a tree without unary nodes and then add the cost introduced by unary nodes. In the first component, one of the kp children incurs $k^{(h-1)\theta}$ from the root and $L_\nu(h-1,k,n)$ for its descendants. Thus, the cost for the kp children of the root is: $kp(k^{(h-1)\theta} + L_\nu(h-1,k,n))$.

Now, consider one of the $k(1-p)$ children of the root without receivers. We again have kp children with connected receivers, and $k(1-p)$ children without connected receivers. A recurrence relation based on this pattern computes the second part of $L_\nu(h,k,n)$ for the $k(1-p)$ children of the root. Consider node σ at level l which does not have receivers connected (refer to Figure 9). There may be receivers at the descendants of σ that use the link from the parent of σ to σ with approximate probability:

$$1 - \left(1 - k^{-l}\frac{k^h - k^l}{k^{h+1} - k}\right)^n \tag{3}$$

where k^{-l} is the probability that a receiver is located below σ, and $\frac{k^h-k^l}{k^{h+1}-k}$ is the probability that the receiver is connected to a non-leaf node at level i, $l < i < h$. The latter probability is based on the fact that the total number of nodes except the root is $k+\cdots+k^h = \frac{k^{h+1}-k}{k-1}$ and the number of nodes at level i is $\frac{k^h-k^l}{k-1}$. We use $1-(1-k^{-l})^n$ as an approximation of Equation (3) for large h values.

Let $T(l)$ denote the cost required to deliver data to the descendants of σ. As illustrated in Figure 9, at least one of the kp children must receive data from nodes other than σ and the descendants of σ. If we consider the additional cost introduced by *unary nodes*, a sibling node of σ which has receivers (π in the figure) minimizes the cost to $2k^{(h-l)\theta} + k^{(h-l-1)\theta}$. An additional cost of $2k^{(h-l-1)\theta}(kp-1)$ is required to relay the data among the kp children of σ. Thus, $B(h-l-1) = k^{(h-l-1)\theta}(2k^\theta + 2kp - 1)(1-(1-k^{-l})^n)$ is incurred for the kp children of σ. Also, $kpL_\nu(h-l-1,k,n)$ is incurred by the descendants of the kp children of σ. For the $k(1-p)$ children of σ without receivers, $k(1-p)T(l+1)$ is incurred. Hence, $T(l)$ can be computed as: $B(h-l-1) + kpL_\nu(h-l-1,k,n) + k(1-p)T(l+1)$. This is equal to $\sum_{i=l}^{h-1} k^{i-l}(1-p)^{i-l} \times \{B(h-i-1) + kpL_\nu(h-i-1,k,n)\}$.

The cost for the $k(1-p)$ children of the root at level $l = 1$ is $k(1-p)T(l=1) = \sum_{i=1}^{h-1} k^i(1-p)^i \times \{B(h-i-1) + kpL_\nu(h-i-1,k,n)\}$.

Therefore,

$$\begin{aligned} L_\nu(h,k,n) &= kp(k^{(h-1)\theta} + L_\nu(h-1,k,n)) \\ &+ \sum_{i=1}^{h-1} k^i(1-p)^i\{B(h-i-1) + kpL_\nu(h-i-1,k,n)\} \end{aligned} \tag{4}$$

Lemma 1 *Solving the recurrence relation in Equation (4) with a fixed ratio $a = \frac{n}{M}$ $(0 < a < \infty)$ (M is as defined in section 4.3) yields:*

$$\begin{aligned} L_\nu(h,k,n) &= k^{(h-1)\theta+1}p + (k^h + k^{h\theta}\sum_{i=2}^{h-1} k^{(1-\theta)i})p^2 \\ &+ k^{(h-2)\theta+1}(1-p)(2k^\theta + 2kp - 1)\sum_{i=0}^{h-2} k^{(1-\theta)i} \\ &- k^{h-\theta}(1-p)(2k^\theta + 2kp - 1)c_2(a,\theta) \\ &+ O(1) \end{aligned} \tag{5}$$

where $c_2(a,\theta) = \sum_{i=0}^{\infty} k^{-(1-\theta)i}e^{-ak^{i+1}}$.

The proof of this lemma and the proof that $L_\nu(h,k,n)$ is the minimum cost overlay tree when receivers are located at any node except the root can be found in [9]. Note that $U_\nu^\theta(h,k)$, the average number of hops on the source to receiver unicast paths, is now computed as: $U_\nu^\theta(h,k) = \frac{1}{M}\sum_{l=1}^{h} k^l \sum_{i=1}^{l} k^{(h-i)\theta}$.

The normalized overlay cost $R_\nu^\theta(h,k,n) = \frac{L_\nu(h,k,n)}{U_\nu^\theta(h,k)}$ does not exhibit a power-law [9]. However, Figure 8(b) demonstrates that $R_\nu^\theta(h,k,n)$ behaves asymptotically similar to a power-law when $0 < a < 1$. In the figure, $R_\nu^\theta(h,k,n) \propto n^{0.83}$. The factor 0.83 is *smaller* than the

0.92 for the case when hosts are only connected at leaves, since many additional hops can be saved in this case. It is also important to note that our decreasing unary node distribution leads to a lower tree cost (0.83 versus an 0.87 factor for this same model with uniformly distributed unary nodes). The cost provides a useful notion for comparing and designing overlay multicast protocols to optimize loads. The 0.8 to 0.9 factor can be also compared to a factor ≈ 0.7 for IP multicast [6, 8].

4.4. Simulation and Experimental Validation

We validate our analytical results using a traceroute-based simulation topology. (Our methodology for synthesizing the routes is discussed in Section 3.1.) We simulate hosts connected to edge routers by randomly connecting 10,000 hosts to the edge routers connected to 60 selected traceroute servers. We first construct an overlay that is a complete graph among these 10,000 hosts. In order to be consistent with our modeling assumption that the least cost overlay tree is used, we compute the minimum spanning tree on that graph. An important difference, however, is that a host in the overlay tree enforces an upper degree bound (UDB) on the maximum number of children, to simulate bandwidth constraints. (Hosts connected to the same router are not considered in the UDB check.)

Figure 8(c) shows the normalized overlay cost versus the number of members with UDB=6. Four different random number generator seeds (RNG_seed=3,5,7,9) are used for the assignment of hosts. We observe that the results are consistent with our modeling results. The normalized overlay cost is asymptotically close to $n^{0.85}$ or so, for a small number of members (< 100). The value was higher ($n^{0.87}$) when we repeated the same experiment with UDB=1. The tree cost saturates at around 36, when the number of members is ≈ 100, which is earlier than the curves in Figure 8(b). This can be attributed to the usage of only 60 routers to which hosts are connected in the simulation, versus 700 to 4000 routers used in Figure 8(b).

We have also examined the normalized overlay cost via simulations of the three overlay protocols on the topologies described in Section 3.2. The results reveal that ESM and MDDBST behave asymptotically close to $n^{0.8} - n^{0.9}$ or so, before they saturate, which is consistent with our analytical results. TAG has a slightly higher cost, due to the u unmatched routers allowed with high $bwthresh$ values. We also found that the normalized cost was higher for the GT-ITM topologies than for the power-law and small-world topologies, since router degree and clustering properties are exploited by overlay protocols to reduce stress and cost.

To further validate our results, we compute the stress and overlay cost for the real ESM tree used in Section 3.1. We find that the maximum stress is 12, the total stress is 696, and the overlay tree cost is 568. Since the average unicast path length is ≈ 12.01, the normalized overlay cost is $\frac{568}{12.01} \approx 47.3$. Since $n = 59$ (we only use hosts for which we could obtain underlying routes), the normalized tree cost $\approx n^{0.945}$.

5. Related Work

Our objectives in this paper overlap with the objectives of work evaluating IP multicast efficiency. Chuang and Sirbu [8] were first to investigate the efficiency of IP multicast in terms of network traffic load. They found that the ratio between the total number of multicast links and the average unicast path length exhibits a power-law with respect to the number of distinct sites with multicast receivers ($m^{0.8}$). Chalmers and Almeroth [6] subsequently investigated the efficiency of IP multicast over unicast experimentally. They argue that the normalized tree cost is closer to $n^{0.7}$ than to $n^{0.8}$. In addition, their results indicate that multicast trees typically include a high frequency (70 to 80%) of unary nodes.

In order to precisely understand the causes of IP multicast traffic reduction, several mathematical models have been devised. Phillips *et al.* [18] were first to derive asymptotic forms for the power-law in k-ary trees and more general networks. Adjih *et al.* [2] obtained more accurate asymptotic forms of the power-law. They abandon the simple k-ary tree used in [18], and use a k-ary *self-similar* tree. However, they provide no experimental data to prove that IP multicast trees are indeed self-similar, i.e., the number of unary nodes decreases as the tree level increases. We consider the case of overlay multicast, not IP multicast, in this paper.

Perhaps the work that comes closest to ours is presented in [19] and [14]. Radoslavov *et al.* [19] characterized real and generated topologies with respect to neighborhood size growth, robustness, and increase in path lengths due to link failure. They briefly analyzed the impact of topology on two heuristic overlay multicast strategies, in terms of stretch and maximum link stress. Jin and Bestavros [14] have shown that both Internet AS-level and router-level graphs exhibit small-world behavior, due to power-law degree distributions and preference to local connections. They outlined how the small-world property affects the overlay multicast tree size.

6. Conclusions and Future Work

We have characterized overlay multicast trees via experimental data and simulations of three overlay multicast protocols. We also have modeled and computed the overlay cost, defined as the total number of hops in all overlay links. Based on our results, we can make the following observations. First, the experimental data and simulations illustrate that both the average delay and the number of hops between parent and child hosts tend to decrease as the level of the host in the overlay tree increases. Our analysis suggests that routing features in overlay multicast protocols, along with power-law and small-world topology characteristics, play a key role in explaining these phenomena. Non-uniform multicast host distribution reinforces them. Second, our models behave asymptotically close to power-laws, ranging from $n^{0.83}$ to $n^{0.92}$ for n hosts. Simulations and experimental data validate our models, and show the latency bandwidth tradeoffs in overlay trees constructed via three different protocols. We can quantify potential bandwidth savings of overlay multicast compared to unicast since $n^{0.9} < n$, and the bandwidth penalty of overlay multicast compared to IP multicast ($n^{0.9} > n^{0.8}$).

One limitation of our experiments is the synthesis of traceroute paths among hosts. Topology inference projects [11] may help us obtain more accurate path information for our future experiments and analysis. We plan to conduct larger-scale simulations and experimental data analysis to better understand overlay tree properties. We will also examine other types overlay protocols, and investigate more dynamic characteristics and performance metrics, including join-leave dynamics, protocol overhead, and delay and bandwidth changes. Finally, we plan to precisely formulate the relationship between the structure of overlay trees, overlay protocols, and Internet topology characteristics. This will ultimately shed more light on overlay protocol design methodologies.

References

[1] Traceroute.org, 2003. http://www.traceroute.org.

[2] C. Adjih, L. Georgiadis, P. Jacquet, and W. Szpankowski. Multicast Tree Structure and the Power Law. In *Proc. of SODA*, 2002.

[3] D. G. Andersen, H. Balakrishnan, M. F. Kaashoek, and R. Morris. Resilient Overlay Networks. In *Proc. of ACM SOSP*, pages 131–145, October 2001.

[4] A. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

[5] M. Castro, P. Druschel, A.-M. Kermarrec, and A. Rowstron. Scribe: A Large-scale and Decentralized Application-level Multicast Infrastructure. *IEEE Journal on Selected Areas in Communications*, 20(8), October 2002.

[6] R. Chalmers and K. Almeroth. Modeling the Branching Characteristics and Efficiency Gains in Global Multicast Trees. In *Proc. of IEEE INFOCOM*, pages 449–458, April 2001.

[7] Y. Chu, S. Rao, S. Seshan, and H. Zhang. Enabling Conferencing Applications on the Internet using an Overlay Multicast Architecture. In *Proc. of ACM SIGCOMM*, pages 55–67, August 2001.

[8] J. Chuang and M. Sirbu. Pricing Multicast Communications: A Cost-Based Approach. In *Proc. of Internet Society INET*, July 1998.

[9] S. Fahmy and M. Kwon. Characterizing Overlay Multicast Networks. Technical Report, August 2003. Available at http://www.cs.purdue.edu/homes/fahmy/.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *Proc. of ACM SIGCOMM*, pages 251–262, August 1999.

[11] P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang. IDMaps: A Global Internet Host Distance Estimation Service. *IEEE/ACM Transactions on Networking*, 9(5):525–540, October 2001.

[12] D. V. Houweling. Internet 2, 2003. http://www.internet2.edu.

[13] J. Jannotti, D. Gifford, K. Johnson, M. Kaashoek, and J. O. Jr. Overcast: Reliable multicasting with an overlay network. In *Proc. of OSDI*, October 2000.

[14] S. Jin and A. Bestavros. Small-World Internet Topologies: Possible Causes and Implications on Scalability of End-System Multicast. Technical Report BUCS-TR-2002-004, Boston University, 2002.

[15] M. Kwon and S. Fahmy. Topology-Aware Overlay Networks for Group Communication. In *Proc. of ACM NOSSDAV*, pages 127–136, May 2002.

[16] N. Malouch, Z. Liu, D. Rubenstein, and S. Sahu. A Graph Theoretic Approach to Bounding Delay in Proxy-Assisted, End-System Multicast. In *Proc. of IWQoS*, May 2002.

[17] L. Peterson, T. Anderson, D. Culler, and T. Roscoe. A Blueprint for Introducing Disruptive Technology into the Internet. In *Proceedings of the HotNets-I*, October 2002.

[18] G. Phillips, S. Shenker, and H. Tangmunarunkit. Scaling of Multicast Trees: Comments on the Chuang-Sirbu scaling law. In *Proc. of ACM SIGCOMM*, pages 41–51, 1999.

[19] P. Radoslavov, H. Tangmunarunkit, H. Yu, R. Govindan, S. Shenker, and D. Estrin. On Characterizing Network Topologies and Analyzing Their Impact on Protocol Design. Technical Report USC-CS-TR-00-731, Dept. of Computer Science, University of Southern California, February 2000.

[20] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker. Topologically-Aware Overlay Construction and Server Selection. In *Proc. of IEEE INFOCOM*, volume 3, pages 1190–1199, June 2002.

[21] S. Shi, J. Turner, and M. Waldvogel. Dimension Server Access Bandwidth and Multicast Routing in Overlay Networks. In *Proc. of ACM NOSSDAV*, pages 83–91, June 2001.

[22] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan. Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications. In *Proc. of ACM SIGCOMM*, pages 149–160, August 2001.