

An Analysis of Multicast Forwarding State Scalability

Tina Wong and Randy Katz
Department of Electrical Engineering and Computer Science
University of California, Berkeley
{twong,randy}@cs.berkeley.edu

Abstract

Scalability of multicast forwarding state is likely to be a major issue facing inter-domain multicast deployment. In this paper, we present a comprehensive analysis of the multicast forwarding state problem. Our goal is to understand the scaling trends of multicast forwarding state in the Internet, and to explore the intuitions that have motivated state reduction research. We conducted simulation experiments on both real and generated network topologies, with a range of parameters driven by multicast application characteristics. We found that the increase in peering among Internet backbone networks has led to more multicast forwarding state at a handful of core domains, but less state in the rest of the domains. We observed that scalability of multicast forwarding state with respect to session size follows a power law. Our findings show that distribution and concentration of multicast forwarding state in the Internet is significantly impacted by application characteristics. We investigated recent proposals on non-branching multicast forwarding state elimination, and found substantial reduction is attainable even with very dense multicast sessions.

1 Introduction

IP Multicast is an efficient point-to-multipoint delivery mechanism because packets disseminated to a large number of hosts travel only once through the common parts of the network. Much research and engineering effort is in progress with the goal of making IP Multicast widely deployed in the Internet. Applications such as large-scale content delivery, software distribution, A/V conferencing, distance learning and network games can all benefit from IP Multicast.

1.1 IP Multicast Routing

The first IP Multicast routing protocol, DVMRP [7], uses a broadcast-and-prune mechanism to create per-source reverse shortest path multicast trees: a source broadcasts packets to the whole network, and leaf routers prune back the distribution when the attached hosts are not interested in the data. CBT [3] and PIM-SM [8] build multiple-source shared trees rooted at a core—both use an explicit-join mechanism, which is more efficient when membership is sparse. An excellent survey of IP Multicast can be found in [2].

Regardless of the underlying multicast routing protocol, a router must maintain membership state to achieve multicast

forwarding. In this paper, we study state in the form of multicast forwarding entries, independent of the specifics of the routing protocol. One such entry determines the distribution of a multicast packet to a router's outgoing interfaces, based on the packet's multicast address, sometimes also the source address, and the incoming interface. Depending on whether a per-source tree or shared tree is used, the amount of memory consumed by a multicast group at a router is either linear or constant to the number of sources.

The number of multicast forwarding entries at a router scales with the number of concurrently active multicast groups in a network. Unlike unicast routing, there is no natural aggregation because hierarchical address allocation and longest prefix match cannot be easily applied. Thus, router memory constraints limit the number of simultaneous multicast groups a network can support. Thaler and Handley state that as the number of multicast groups increases, scalability of multicast forwarding state is likely to be the biggest issue facing multicast deployment in the future [19].

1.2 Multicast State Reduction

Next, we describe recent works that propose reducing multicast forwarding state (thereafter, state or multicast state) through aggregation, tunneling, non-branching state elimination and application-level clustering.

State aggregation

Forwarding state aggregation replaces multiple forwarding entries with a single one, if the entries have adjacent group address prefixes and matching incoming and outgoing interfaces. It has been shown that hierarchical address allocation and membership locality increase the degree of state aggregation [19]. Leaky aggregation [15] trades bandwidth for state, and allows data to "leak" downstream to links without receivers.

Tunneling

One common observation is that multicast state scalability is most critical at the "core" or "backbone" routers, because intuitively, most multicast trees pass through these highly connected routers to span their members. The label stacking mechanism in MPLS [5] can set up DVMRP-style tunnels to eliminate multicast state at the intervening routers. Distributed Core Multicast (DCM) [4] also takes an analogous approach: it places multiple specialized agents at the edge of the backbone network and configures tunnels among these agents, so multicast state is only kept at the ingress and

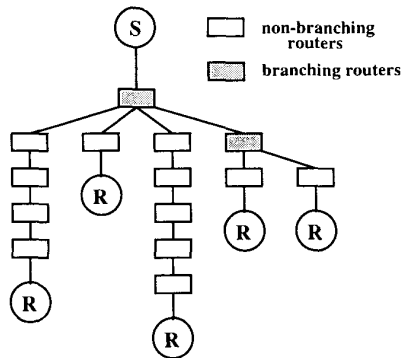


Figure 1. An illustration of non-branching multicast state. A multicast tree is shown connecting 1 source and 5 receivers. The shaded nodes are branching routers, whereas the white ones are non-branching.

egress routers.

Non-branching state elimination

Another observation is that multicast trees for sparse groups, those with small to medium number of receivers, are likely to have chains of routers that forward multicast packets to only one outgoing interface [20]. Dynamic Tunnel Multicast (DTM) [20] and REUNITE [18] enhance multicast state scalability through eliminating such non-branching state. Figure 1 is an illustration: only the branching routers need to carry out multicast forwarding, and keeping membership state at the non-branching routers is unnecessary.

Application-level clustering

Large-scale applications such as content delivery and distributed interactive simulation (DIS) use a significant number of multicast addresses. Clustering [11, 23, 24] exploits application-level knowledge, such as users' preferences in content or players' positions in virtual space, to intelligently group sources and receivers into a limited number of multicast groups.

1.3 Our Contributions

Although much research has focused on the reduction of multicast state, we know of no comprehensive study that analyzes the problem of multicast state scalability itself. The contribution of this paper is to provide that analysis. We investigate the following questions and summarize our results briefly here. The results are applicable to future network provisioning, and useful in multicast protocol and application design.

Scaling trends

How do multiplying peering agreements among parallel backbone networks affect multicast state? We found that increasing peering in the Internet leads to higher multicast memory requirements at a handful of well-connected core domains, but lower requirements at the rest of the domains,

when all other factors remain equal.

How do rising subscriptions to individual applications affect multicast state? We observed that the growth of multicast state with respect to session size follows a power law of the form $y \propto x^a$, where y is the fraction of routers or domains that are stateful, x is group size, and a is a constant. Furthermore, the exponent a has remained relatively constant for the Internet over the past 3 years.

State distribution

How concentrated is multicast state at the "core" routers? We found that state distribution and concentration is rather sensitive to application characteristics such as the number of members and their locations in the network. We concluded that tunneling can effectively reduce multicast state only when membership is extremely sparse and spread-out.

Potential state reduction

Are non-branching routers stateful? Are we fixing the multicast state problem at the wrong routers? Our results show that significant state reduction is possible with non-branching state elimination even with dense sessions, making it a promising approach to tackle the multicast state scalability problem.

1.4 Paper Roadmap

The rest of the paper is organized as follows. §2 discusses the methodology we used in studying the multicast state problem. §3 examines how topological properties affect multicast state, and the scaling trends in the Internet over the last 3 years. §4 looks at the distribution and concentration of multicast state, and the effects of group member locations in the network. §5 studies the scalability of multicast state with respect to session size, and shows how it follows the power law. §6 quantifies the merits of non-branching state elimination schemes. We discuss related work in §7 and conclude the paper in §8.

2 Methodology

We conducted simulation experiments in this analysis. To achieve our goal of a comprehensive study, we varied 3 parameters in the simulation—topology, session density and membership model, and measured 2 metrics—local state and true local state. In this section, we also discuss the assumptions we made and the potential shortcomings they might cause.

2.1 Topology

Table 1 lists the topologies we used in the experiments. The AS graphs correspond to connectivity among Internet autonomous systems (AS), such that each node in a graph represents an AS. These graphs are available from the National Laboratory for Applied Network Research [13]. To understand state scalability trends, and to see whether our observations are time invariant or at least somewhat predictive of the near future, we selected 4 AS graphs that span approximately

Name	Type	Date	# Nodes	# Edges
transit-stub	gen	NA	1000	1836
TIERS	gen	NA	5000	7089
MBone	real	02/23/99	4179	8568
AS-Nov97	real	11/28/97	3100	5559
AS-Jun98	real	06/15/98	3736	7140
AS-Mar99	real	03/24/99	4830	9078
AS-Jan00	real	01/02/00	6474	13260

Table 1. Topologies used in this paper.

the last 3 years. The earliest graph was collected in November 1997; the most recent in January 2000. The MBone graph was collected by the SCAN project at USC/ISI [16] in February 1999, and each node represents a MBone router. We also included 2 generated topologies in our experiments. The GT-ITM package [10] creates transit-stub style topologies. A transit domain is multi-homed and connects multiple single-homed stub domains and other transit domains together. The TIERS [21] generator categorizes routers into LAN, MAN and WAN routers when constructing a topology. In the rest of the paper, we use the term “node” to denote either a domain or router in a topology.

2.2 Session Density

Intuitively, the amount of multicast state in a network is dependent on the number of members in the multicast groups. Since topology sizes range from 1000 to 6474 nodes, we use session density instead of absolute session size to describe the number of nodes in a group. There are 3 ranges of session density in the experiments: sparse, from 0.1% to 0.9% of topology size; medium, ranging from 1% to 9%; and dense, from 10% to 90%. For example, a multicast session of 1% density in a topology with 2500 nodes has 2500×0.01 or 25 members. Throughout the paper, we use θ to denote session density.

2.3 Membership Model

Another factor that affects multicast state is the location of multicast group members in the network. We model a spectrum of membership types with a membership taxonomy¹, illustrated in Figure 2. The taxonomy is divided into 2 axes—topological correlation within a single multicast session, and subscription correlation across multiple sessions—from which we define 6 models. Even though this taxonomy aims to cover all possible membership models, it is conceivable that future multicast applications will not fit into it.

Using trace data from actual multicast sessions would be more realistic, but IP Multicast is not widely deployed and its usage is limited to mostly research and testing. Studies have shown that most of the MBone’s activity is among a relatively small group of users [1].

We describe the 6 models next; the number in the parentheses corresponds to those in the taxonomy.

Random (1)

¹We would like to acknowledge Mark Handley on his role in defining this taxonomy.

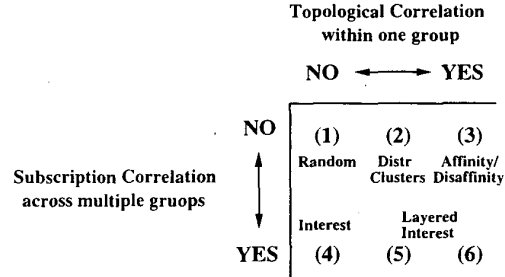


Figure 2. Membership taxonomy. A membership model is defined based on whether there is topological correlation within a single multicast session, and/or subscription correlation across multiple sessions.

This models applications with sources and receivers randomly located in the network, such as large-scale content delivery, software upgrades and stock quotes dissemination.

Affinity/disaffinity (3)

We use the affinity model to emulate applications with members that tend to cluster together, e.g. local news and traffic reports. The disaffinity model is used to capture applications with members that tend to spread out, like video conferencing and electronic whiteboards. In the simulation, we implemented affinity and disaffinity after [14], as follows. Suppose we are creating a session g of m nodes in topology t . The first node is selected randomly from t . To pick the k th node, we assign a probability p_i to each node $n_i \notin g$, defined as $p_i = \frac{\alpha}{w_i \beta}$, where $w_i = \min_{n_j \in g} d(n_i, n_j)$, where $d(n_i, n_j)$ is the distance in hop counts between n_i and a node n_j already in g , and α is calculated such that $\sum_{n_i \notin g} p_i = 1$. We use $\beta = 15$ and $\beta = -15$ for affinity and disaffinity, respectively, to model the extreme case when members are as close together or as spread-out as possible.

Distributed clusters (2)

In this model, members in a multicast session are divided into a few clusters scattered randomly in the topology, and within each cluster the members close together [19]. Example applications include distance learning, training videos and lectures, in which there are regions of receivers interested in the multicast data. We implemented distributed clusters in the simulation by randomly selecting 5 nodes in a topology as the first member of 5 evenly sized clusters, and within each cluster using the affinity model to pick the rest of the members.

Interest (4)

We use the interest model to emulate overlapped user subscriptions across multicast applications. For example, a portion of users subscribed to the stock quotes multicast channel are probably also listening to the business news channel, and membership on each channel is scattered over the globe. In the simulation, we implemented interest as follows. Suppose we are creating sessions of m nodes. The first group g_1 is selected randomly. For the k th group g_k ,

first a fraction δ of the m nodes are picked randomly from $n \in g_{(k-1)}$ (nodes in the previous group) then $(1 - \delta)$ of the m nodes are chosen from $n \notin g_k$ (nodes not already in this group). We examined $\delta = 0.1, 0.5$ and 0.9 , or 10%, 50% and 90% interest correlation, in the experiments. The results for the 3 factors are quite similar, and we only present 50% in this paper.

Non-random interest (5-6)

In non-random interest, member locations within a multicast group follow the affinity/disaffinity or distributed clusters model, as well as exhibiting correlation across sessions. Example applications include layered multicast and distributed interactive simulations, which use multiple multicast addresses. These applications are the most far-off, and we only examine membership models (1)-(4) in this paper.

2.4 Experiments

In each experiment, we first fixed the session density θ , topology t , and membership model l . Then, we picked a set of nodes from t according to θ and l , and built a shortest path tree based on hop counts rooted at a random node from this set. This models a single source-based tree or a core-based tree with multiple sources rooted at one of the sources. We repeated the experiment 1000 times, after which we calculated local state and true local state at each node with equations (1) and (3), respectively. We explain these 2 metrics later. We used all combinations of θ , t and l , where each combination forms a set of experiments, yielding $|\Theta| \times |T| \times |L|$ or $27 \times 7 \times 5 = 945$ different sets of results.

There are 3 caveats in our study. First, our topologies are symmetric, whereas routes might be asymmetric in the real world. Second, because we do not have available to us inter-domain BGP edge costs, which are determined by policy-based routing metrics, we use hop counts instead when building shortest path trees. Third, we choose a member of a group as the root or core of a multicast tree, which is not always true for shared tree construction. Optimal core location is NP-complete thus impractical for the Internet [22]; currently, heuristics such as closeness to high bandwidth sources are used to choose a core.

2.5 Local State

After each set of experiments, we measured the amount of multicast *local state* s_i at each node n_i in a topology, which is the fraction of the simultaneous multicast groups n_i has to carry. To state mathematically,

$$s_i = \frac{p_i}{r} \quad (1)$$

where p_i is the number of groups that passed through n_i , and r is the number of simultaneous multicast groups, i.e. the number of experiments or 1000². Throughout the paper, we report this metric in %: when $s_i = 0\%$, none of the r groups goes through n_i ; when $s_i = 100\%$, n_i stores multicast state for all r groups. Recall that the AS graphs represent the Internet at the domain-level, and we measured local state for the

²To simplify the analysis, this metric measures state in the number of multicast forwarding entry. In reality, the amount of memory used by an entry varies, depending on the number of interfaces at the router.

border routers within each domain. The local state at each node of an AS graph is actually the union of local state for all border routers within the associated domain. This is the maximum local state a border router might store if the router is involved in all of the multicast groups. In reality, local state at each border router might be lower—there might be several border routers within each domain, presumably more when the domain spans a wider region. A multicast group generally passes through a few and not all border routers in a domain. Thus, the more border routers in each domain, the less local state each border router maintains, because local state is more distributed or spread out across the routers.³

Since the experiments in each set are independent, except for the interest model, we can apply the central limit theorem to compute the 95% confidence intervals of local state. We found that $r = 1000$ is sufficient to yield a tight confidence bound.

We define *global* or *total* state in a topology as the sum of local state at each node,

$$\sum_{n_i \in T} = s_i \quad (2)$$

2.6 True Local State

We also measured the reduction of local state possible from non-branching state elimination schemes such as DTM, REUNITE and MPLS. A node n_i stores non-branching state for a multicast group g if n_i forwards data only on 1 outgoing interface for g , and such state can be eliminated because multicast forwarding is unnecessary. After each set of experiments, we first calculated the non-branching fraction f_i at each node n_i as,

$$f_i = \frac{p_i'}{p_i}$$

where p_i is the number of groups that passed through n_i , and p_i' is the number of those groups that was non-branching. If $f_i = 0$, all groups that n_i stores state for are branching; if $f_i = 1$, all groups are non-branching. We use the non-branching fraction at a node to compute its *true local state* s_i' , i.e. s_i' is the amount of local state n_i maintains for only the branching groups when non-branching state is eliminated. To state mathematically,

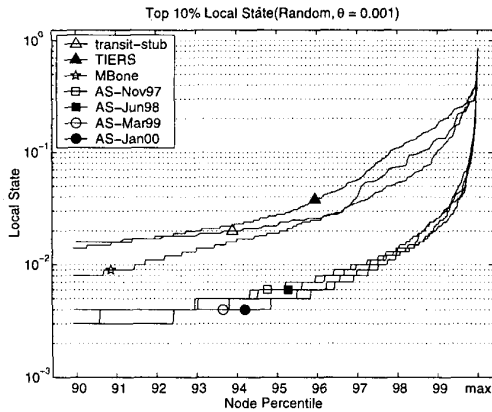
$$s_i' = s_i(1 - f_i) \quad (3)$$

where s_i is the original local state at n_i .

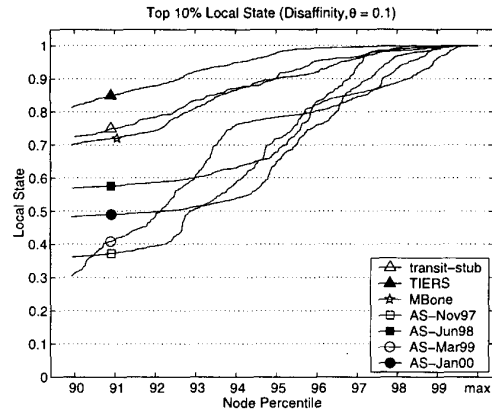
3 State Scaling Trends with respect to Topological Properties

This section discusses multicast state scaling trends in the Internet. Our results show that topological properties such as connectivity affect amount of state. To explain this, we first examine the differences in state among the 7 topologies. Figure 3(a) plots local state at the top 10% nodes with

³We present a simple calculation of local state at each border router in a domain. Let d be the degree of the domain thus the number of border routers. Assume each multicast group passes through 2 border routers as part of transit. Then, on average local state $s_{i,j}$ at a border router j of domain i is $\frac{2s_i}{d}$, where s_i is defined by equation 1.

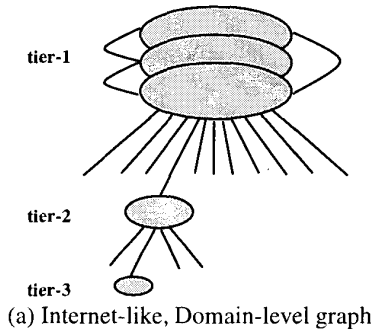


(a) Random membership, 0.1% session density

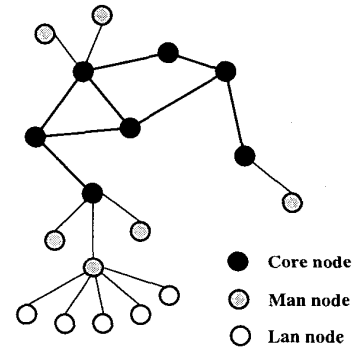


(b) Disaffinity membership, 10% session density.

Figure 3. *Effects of topology.* Differences in the amount of local state at the top 10% nodes most stateful nodes. The 4 AS topologies have similar results, and have fewer state than the MBone and generated topologies at most of the nodes.



(a) Internet-like, Domain-level graph



(b) MBone-like, Router-level graph

Figure 4. *Differences in topology.* In the inter-domain level (a), it takes relatively few hops to go from one node to another in the graph. There are several parallel “tier-1” nodes that are extremely well-connected. Although a hierarchy is also observed in (b), the outdegree of the core nodes are much lower. It takes more hops to reach a node than in (a).

the most local state for each of the 7 topologies, when random membership is used to create sessions of 0.1% density. Local state at a node n_i is the fraction of concurrent multicast groups that n_i maintains state for, computed using equation 1. For each topology, we sorted the nodes in increasing order of local state and plotted node rank against local state. From 90-percentile to roughly 99-percentile, the non-AS topologies—transit-stub, TIERS and MBone—have up to five factors more local state than the AS topologies. We also observe the same behavior for nodes below 90-percentile, but focus on the busiest nodes here. The 95-percentile nodes in transit-stub, TIERS and MBone keep about 2.4%, 2.7% and 2.2% local state, respectively; the 4 AS topologies store only about 0.5% to 0.6%. However, at the few most stateful nodes, the AS topologies have up to two factors more state than the non-AS topologies: the 100-percentile nodes in transit-stub, TIERS and MBone have roughly 55%, 46% and 38% local state, respectively, whereas the 4 AS topologies maintain 58%, 71%, 70% and 86%, in chronological order. Notice that the results for the 4 AS topologies are very similar, except for

the few most stateful nodes. We also found similar local state differences among the topologies with a range of session densities and all membership models except affinity. Figure 3(b) plots the same thing as Figure 3(a), except disaffinity is used to generate sessions of 10% density.

3.1 Hypothesis

Why do the topologies have different amount of state? Let us examine some properties of the topologies. The Internet is divided into *tiers*. Tier-1 networks are Network Service Providers (NSPs) or carriers such as UUnet and MCI, which transit traffic with other NSPs according to some peering agreements. Tier-2 networks are Internet Service Providers (ISPs) such as Concentric and Earthlink, which construct their networks from long-haul capacity from a few NSPs. Tier-3 networks are smaller, regional corporate or campus LANs. The definition of tiers is becoming fuzzy, as major corporations can peer directly with a carrier to get high bandwidth access, for example. Figure 4(a) illustrates such a hier-

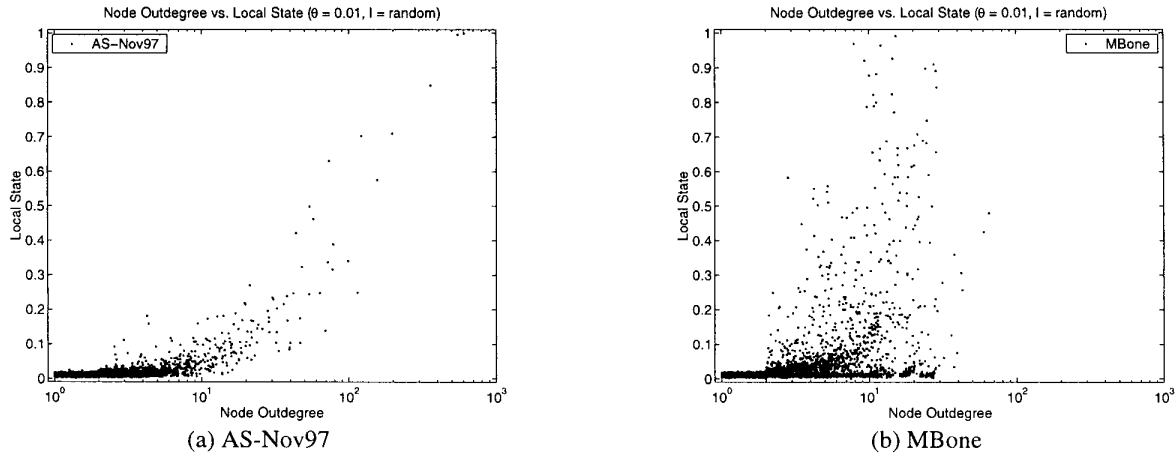


Figure 5. Node local state vs. node outdegree. Random membership is used to create sessions of 1% density for both topologies. Nodes with larger outdegrees keep more local state.

archical structure. The MBone, on the other hand, is a virtual overlay network on top of the physical Internet, with DVMRP tunnels connecting backbone routers to provide long-distance multicast connectivity. The backbone routers connect to local routers within their regions which wish to send and receive multicast data, and the local routers further to their local routers, and so on. The connectivity is ad-hoc and manually configured, which means the MBone is less structured and more poorly connected than the AS graphs. Figure 4(b) shows an MBone-like topology.

Topological properties of a graph directly affect the fanout and height of multicast trees built on top of it, which in turn determine the amount of multicast state maintained by the nodes in the graph. From one viewpoint, a multicast tree constructed on a highly connected graph would have a large fanout and short height. A few core nodes supporting the highest degrees of connectivity are involved in most of the multicast trees, allowing the rest of the nodes (even the core ones) to be less stateful. From another perspective, a multicast tree built on a poorly connected graph would be deeper with a smaller fanout. More nodes need to share the responsibility of multicast forwarding, attaining a load balancing effect.

In recent years, peering has increased tremendously in the Internet backbone, which is one intuition of why the AS topologies have less local state for a significant portion of their nodes than the non-AS topologies, but more local state for a few "core" nodes.

3.2 Confirmation

To confirm this hypothesis, Figures 5(a) and (b) are scatter plots of node local state as a function of node outdegree, when random membership is used to create sessions of 1% density in the AS-Nov97 and MBone topologies, respectively. Node outdegrees directly relate to connectivity of the graph. Note the log scale on the x-axis. Each point (x_i, y_i) in the figure represent a node n_i in the topology, where x_i is the outdegree or number of immediate neighbors at n_i and y_i is the local state of n_i . Random jitter is added to each point to better dif-

ferentiate the points. In general, nodes with larger outdegrees keep more local state. Most nodes in AS have considerably less local state than those in the MBone, as we have already seen from Figures 3(a)-(b). Moreover, a few nodes in AS have much larger outdegrees than those in the MBone.

To probe further, we examine the connectivity of the topologies by plotting their cumulative distributions of node to outdegree. The distributions are almost identical for the 4 AS topologies⁴, whereas they are quite different from the non-AS topologies⁵. Approximately 80% to 85% of nodes in the AS topologies have outdegrees 3 or less, whereas there are only about 50% to 73% in the non-AS topologies. More importantly, the distributions for the 4 AS topologies have much longer tails, with maximum node outdegrees of 605, 817, 1094 and 1459, in chronological order. The transit-stub, TIERS and MBone topologies have maximum node outdegrees of only 12, 30 and 64, respectively.

We also plot the cumulative distributions of all-pairs shortest path to path length in hop counts for all 7 topologies in Figure 6. Path lengths directly relate to heights of multicast trees built on top of the graphs. Again, the distributions for the 4 AS topologies are almost identical, but strikingly different from the ones of non-AS topologies. About 80% of the all-pairs shortest paths in the 4 AS topologies have 4 hops or less, which is true for only 5% in the non-AS topologies. We also examine the shortest path lengths from the top 10% nodes with the most local state, and found the same pattern of differences.

⁴This agrees with the findings in Faloutsos et al. [9], in which the authors found that the frequency f_d of an outdegree d is proportional to the outdegree to the power of O , and O is practically constant for 3 AS topologies between November 1997 and December 1998. In other words, the node outdegree distribution follows the power law of the form $f_d \propto d^O$.

⁵Using the method described in [9], we calculated the outdegree exponents and ACC pairs for the MBone, transit-stub and TIERS to be (-1.67, 0.93), (-1.84, 0.77) and (-2.18, 0.96) respectively. Only the exponents for the TIERS topology matches the ones found for the AS topologies.

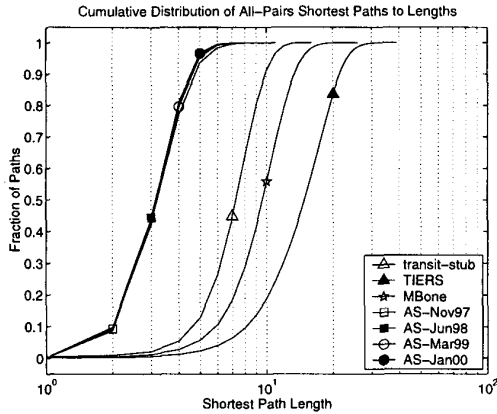


Figure 6. Cumulative distribution of all-pairs shortest path lengths. The MBone and generated topologies have substantially longer path lengths than the 4 AS topologies, which have almost identical distributions.

θ /state	90-percentile	95-percentile	99-percentile
0.1%	0.4 → 0.3	0.6 → 0.5	2.1 → 2.3
1%	2.6 → 2.1	5.0 → 3.8	18.0 → 19.6
10%	21.5 → 18.9	37.7 → 31.0	83.0 → 84.9

Table 2. Local state variations in AS topology between November 1997 and January 2000. Random membership is used to create sessions of 0.1%, 1% or 10% density. Local state is in percent. The amount of local state decreases for the 90 and 95-percentile nodes, but increases for the 99-percentile nodes, for all 3 session densities.

3.3 Internet Trends

Inter-domain BGP peering results in shorter path lengths, which is desirable, but comes at a cost of increased state thereby raising the memory and processing requirements at a handful of core domains. However, this cost is offset by reduced state at the rest of the domain. An interesting implication arises: as the Internet evolves with increasing peering, the memory requirements at almost all of the border routers actually decline, while the most highly connected routers need to maintain multicast state for almost all concurrent groups, assuming the number of concurrent multicast groups remains constant. We observe this in our results. There is a slight trend of local state decreasing between the 90-percentile to 99-percentile ranked nodes, but increasing from 99-percentile and up, over this time period. This holds true for a range of session densities and most membership models. Table 2 lists the decline in local state at the 90-percentile and 95-percentile nodes, and the rise at the 99-percentile nodes. For example, at 1% session density, the 90-percentile node in AS-Nov97 holds 2.6% local state, whereas in AS-Jan00 it holds only 2.1%.

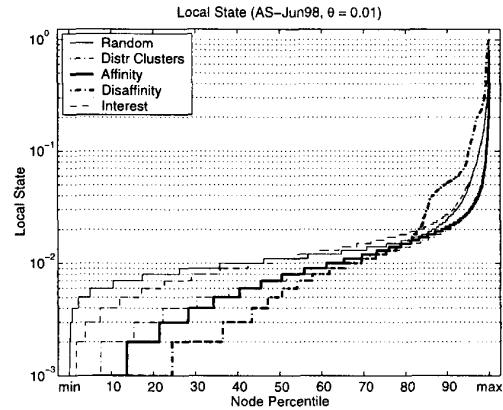


Figure 7. Distribution of local state with respect to membership: Sparse sessions. The session density is 1%. The 4 AS topologies yield almost identical results. The other topologies also have similar skewed distributions, but the membership models do not converge at the highest node rank. Namely, with affinity, the amount of state at the node with the most local state is about an order of magnitude lower than other membership models.

4 State Distribution and Concentration

In the previous section, we examined state scaling trends with respect to topology. Here, we discuss the general distribution of multicast state across a topology as a whole, and study how session density and membership impact this distribution.

4.1 Sparse Sessions

Figure 7 plots local state as a function of node rank for the AS-Jun98 topology, when session density is 1% for all 5 membership models. The other 6 topologies exhibit similarly skewed distributions, and we omit their figures because of space. As we observed previously, state is concentrated—a handful of nodes have much local state and the rest very little. We calculated the percentage of global state that the 10% nodes with the most local state hold. Examining the degree of state concentration gives us insight on the statement, “state is mostly in the core for small groups”. Our results show that the maximum state concentration is 87%, when disaffinity is used to create sessions of 0.1% density in the MBone topology. In other words, 10% of nodes in the MBone hold 87% of the total state. Other membership models demonstrate lower state concentration: roughly 70% to 87% for disaffinity; from 60% to 70% for random; only 30% to 50% for affinity.

We see from above that multicast state is only highly concentrated, thus tunneling mechanisms are only beneficial, when memberships are sparse and spread-out. Applications that meet these criteria are limited. One example is video conferencing or collaborative editing, which usually has a small number of participants separated by long-distances, where it is inconvenient to conduct face-to-face meetings. Network

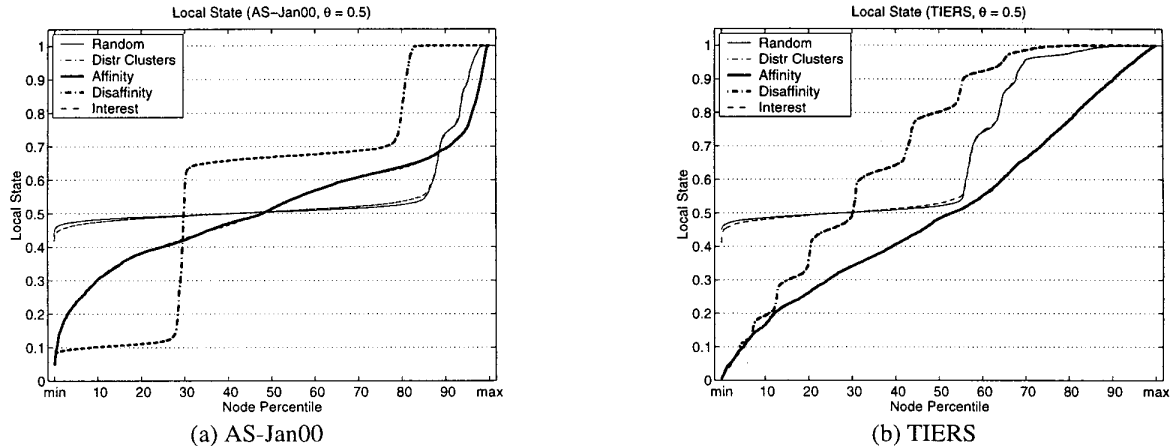


Figure 8. Amount and distribution of local state with respect to membership: Dense sessions. The session density is 50%. The 4 AS topologies all exhibit the 3-level “steps” effect in the distributions. The non-AS topologies have about 6-7 levels. Disaffinity still has the largest amount of state, especially at the most heavily loaded nodes. Random and interest membership models show almost the same results at this session density. Likewise for affinity and distributed clusters.

games might not fit into this category, since bandwidth and latency constraints often bring well-connected players together to attain a quality gaming experience.

4.2 Effects of Member Clustering

We also observe in Figure 7 that from about 80-percentile and up, disaffinity membership has the busiest nodes and affinity the least. Namely, the 90-percentile node keeps 5.2%, 2.7%, 2.4%, 2.1% and 2% local state, for disaffinity, interest, random, distributed clusters and affinity, respectively. This membership ordering corresponds to decreasing levels of member clustering. We also found this same ordering in the non-AS topologies as well, although the local state is higher, as described in the last section. Why does disaffinity yield the busiest nodes and most state concentration, and affinity the least? Intuitively, as members are spread further apart, tree size increases since more nodes are on the tree, thus local state at each node rises. By the same token, as member clustering increases, tree size decreases because fewer nodes are on the tree, therefore the local state at each node reduces. This agrees with the findings in Phillips et al. [14], which show that member clustering significantly affects multicast tree size.

On the other hand, at nodes of roughly 80-percentile and below, disaffinity has the least local state of all membership models. For example, the 50-percentile node has 0.5%, 1.1%, 1.1%, 0.7% and 0.7% local state, for disaffinity, interest, random, distributed clusters and affinity, respectively. This is an artifact of how disaffinity is defined: a node on a multicast tree built using disaffinity is either a poorly connected node, chosen because it is farthest away from other group members; or the node is a well connected node, on the shortest path from the root of the tree to some member(s) in the group. That is, a poorly connected node is unlikely to be on a multicast tree unless the node is selected as a member of the multicast group. With random membership, there is equal probability for each node to be selected as a group member, which explains the

highest local state at the lower ranking nodes. Analogously, it is probable for a poorly connected node to be on a multicast tree when membership follows affinity, because group members are close.

4.3 Dense Sessions

Figures 8(a) and 8(b) plot local state versus node rank for the AS-Jan00 and TIERS topologies when each of the 5 membership models is used to create sessions of 50% density. Almost all nodes have 10% or higher local state. As expected, state is more evenly distributed in the topology here—the top 10% nodes with the most local state hold only about 15% to 20% of the total state—for all combinations of topology and membership. We see obvious “steps” in the distributions for disaffinity membership that do not manifest in other models. There are 3 “levels” or “plateaus” for the 4 AS topologies, and about 6-7 for the non-AS topologies. The “steps” arise from the hierarchical nature of the graphs. As described in Section 3, the Internet is comprised of about 3 tiers, which conforms to the 3 “levels” in Figure 8(a). When members within multicast groups are spread far apart as in the disaffinity model, the tier-1 nodes are on almost all the multicast trees, the tier-2 nodes are relatively less likely to be on them, and the tier-3 nodes the least probable. We have also found that connectivity in the Mbone and the generated topologies is poorer, in terms of lower node outdegree and longer path lengths, thus causing narrower “plateaus” or more “levels” in Figure 8(b).

5 State Scaling Trends with respect to Session Density

This section investigates the growth rate of multicast state across the Internet as the popularity of individual multicast applications rises. As we found in the previous section, multicast state can be quite concentrated, especially when sessions are sparse. Coming up with a meaningful metric to summarize local state over all nodes in a topology is tricky. Taking

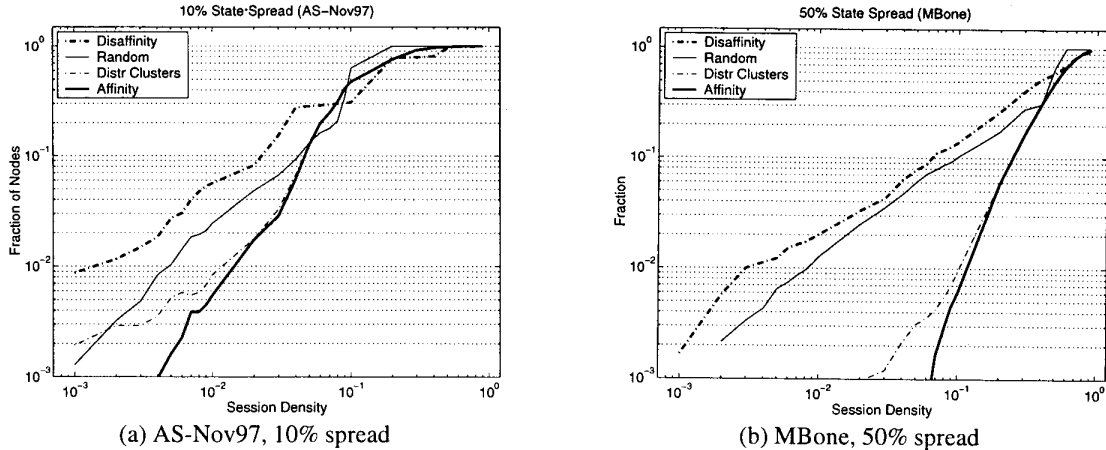


Figure 9. Scalability of multicast state with respect to session density. State spread grows proportional to some constant power of session density, which follows a power law. Note the log-log scale.

Topology	Random	Distr Clusters	Affinity	Disaffin
transit-stub	0.86	0.87	1.23	0.62
TIERS	0.74	0.74	1.47	0.50
MBone	0.78	0.88	1.53	0.60
AS-Nov97	1.12	1.16	1.36	0.76
AS-Jun98	1.16	1.18	1.36	0.79
AS-Mar99	1.09	1.20	1.37	0.78
AS-Jan00	1.10	1.25	1.42	0.80

Table 3. Power law exponents. The 4 AS topologies have very similar exponents within each membership model, and relatively larger exponents than the ones found in the non-AS topologies.

the mean and standard deviation is inappropriate, as we are concerned mainly with the most stateful nodes. We define $x\%$ spread, which is the fraction of nodes in a topology with $x\%$ or more local state. For example, if half of the nodes have 10% or more local state, then the 10% spread is $\frac{1}{2}$. Figure 9(a) plots the 10% spread versus session density in AS-Nov97 on a log-log scale, for all membership except interest, which is omitted as it is almost identical to the results of the random model. Figure 9(b) plots the 50% spread in MBone.

As expected, spread increases with session density. We applied linear regression on the $(\log(\theta), \log(y))$ pairs for each combination of topology and membership, and validated the accuracy of the regression with the correlation coefficient. All coefficients fall in the range between 0.9 and 0.99; 1 denotes perfect correlation. The strong correlation suggests that the growth of multicast state is proportional to some constant power a of session density θ , i.e., $y \propto \theta^a$ where y is spread, fitting in the form of a power law [9]. The values of the exponent a for 10% spread are listed in Table 3—the larger the exponent, the faster multicast state spread grows. The growth of state spread is inversely proportional to member clustering. For all 7 topologies, the following ordering holds: affinity

has the largest power law exponents, followed by distributed clusters, then random, and disaffinity shows the smallest exponents. The exponents range from 0.50 to 1.53. This variation is explained by the fact that the different membership models exhibit different spread when sessions are sparse, but the spread converges to 100% for all models as sessions become dense. For example, at 1% session density, the 10% spread for affinity is roughly 0.005 of the topology, but for disaffinity it is at about 0.06—more than an order of magnitude higher. At 50% session density, however, the 10% spread for all membership is 1: all nodes in the topology.

5.1 Applicability of Power Law

Notice that the 10% spread exponents with each membership model are quite similar for the Internet over the past 3 years. The exponents for random range from 1.09 to 1.16; for distributed clusters 1.16 to 1.25; for affinity 1.36 to 1.42; for disaffinity 0.76 to 0.80. If we assume the basic topological structure in the Internet remains constant [9, 12], then we expect the exponents would stay similar. We can apply the multicast state power law to roughly approximate spread growth for the Internet in the near future. In particular, it is possible to estimate the increase in stateful routers or domains as subscriptions of certain multicast applications expand. For example, let us assume market research finds that the subscription levels of all Internet TV multicast are to double from 5% to 10%, and currently 30% of the routers or domains need to maintain state for 10% or more concurrent multicast groups. If users of such applications are randomly located in the Internet, then the future state spread y_f is estimated as,

$$y_f = \frac{y_c}{\theta_c^a} \times \theta_f^a$$

where y_c is the current state spread (30%), θ_c is the current subscription density (5%), θ_f is the future density (10%), and a is the exponent for random membership (1.10). The new state spread is roughly 64%, a little more than double.

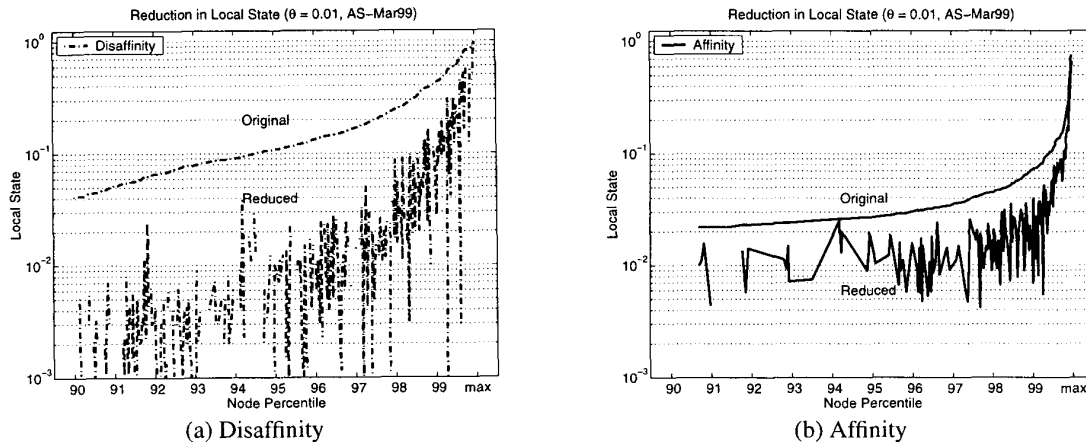


Figure 10. Reduction in local state attainable through eliminating non-branching state. The more member clustering, the less reduction is possible. The results for AS-Mar99 with 1% session density is shown.

6 Potential State Reduction

This section examines the reduction in multicast state attainable with non-branching state elimination schemes. Particularly, we studied the degree of reduction at nodes with the most local state and when session density is high. Our results show that significant improvement is possible for both cases. Measuring the overhead of signaling control and protocol complexity of non-branching state elimination schemes is left as future work.

Figure 10(a) plots original local state and true local state versus node rank, in AS-Mar99 when disaffinity is used to create sessions of 1% density. The top line depicts the original local state for the top 10% nodes with the most local state, and the bottom line is the unsorted true local state for the same nodes. True local state at each node is computed using equation (3), which represents the state used to store just the branching groups. In general, non-branching state is quite prevalent at 1% session density. Up to two orders of magnitude in state reduction is achieved here: at the 95-percentile node, local state drops from approximately 12.4% to 0.1%. Moreover, some nodes see complete multicast state elimination, where all multicast groups passing through these nodes are non-branching. The improvements are smaller for nodes with the most local state: a few factors of reduction are possible, and a negligible fraction of nodes see zero improvement.

Figure 10(b) plots the same thing, but when affinity is used. The degrees of reduction are substantially lower here, but nonetheless up to an order of magnitude is achieved. It follows intuition that the more member clustering within a group, the fewer non-branching nodes on the corresponding multicast tree, because it is less likely the tree has long chains spanning across the topology.

Even when session density is as high as 90%, non-branching state remains prevalent. Membership does not make much impact on the degree of reduction. The intuition is multicast groups passing through poorly connected nodes are almost always non-branching, regardless of session density. Since almost all nodes have high local state at 90% session density, the reduction factors at the poorly connected

nodes are quite substantial, because a major portion of the state can be eliminated.

7 Related Work

In this section, we discuss related work in Internet topology discovery, multicast cost analysis, and multicast tree construction comparisons. These results have greatly influenced our work.

Faloutsos et al. [9] discovered statistical regularities in the Internet that can be captured precisely in the form of power laws. The power laws characterize highly skewed topological properties such as node outdegree, rank and neighborhood size. Moreover, the exponents of the power laws remain constant for 3 Internet instances over a 1 year time-span. We found that the scalability of multicast state with respect to session density also follows a power law, and that the exponents for the Internet over the past 3 years are again relatively similar. Medina et al. [12] conducted a follow-on study to investigate the possible causes of power law in the Internet. The authors found that preferential attachment and incremental growth are the key contributors. Specifying node outdegrees and ranks are essential to generate topologies that closely resemble that of the Internet.

Chuang and Sirbu [6] detected another power law—multicast link usage over unicast is accurately predicted with membership size to the constant power of 0.8, and holds for a range of generated and real network topologies. In other words, the cost of a multicast tree connecting m members is proportional to $m^{0.8}$. The authors advocate pricing multicast according to membership size with a two-part tariff structure. Phillips et al. [14] followed up on this study, and found that multicast link usage scales linearly with membership size with a logarithmic correction for networks with exponential reachability. The authors also found that member clustering significantly affects the multicast tree size, which agrees with our results.

Wei and Estrin [22] investigated the trade-offs between shared trees and source-specific shortest path trees, in terms of end-to-end delay, bandwidth consumption and traffic con-

centration in the network. The authors found that there is no universal type of tree that is suitable for all application types. Shared trees are more efficient from an individual multicast group's point-of-view, but can create hot spots in the network. Source-specific trees, on the other hand, exhibit less delay and more evenly distributed traffic concentration, but bandwidth consumption is higher. The cost of tree state is left as future work.

8 Summary

In this paper, we have presented a comprehensive study on multicast state scalability. We have conducted simulation experiments on both real and generated network graphs, with a spectrum of parameters driven by multicast application characteristics. One of the parameters is membership—which we defined using a taxonomy encompassing a spectrum of possible models. This also can be applied to other multicast related analyses. Our results are summarized as the following points:

- Increased network peering results in shorter path lengths, which is desirable but comes at a cost of more multicast state at a handful of core routers or domains, but is offset by reduced state in a majority of routers or domains.
- Scalability of multicast state with respect to session density follows a power law, i.e. the fraction of routers or domains that are stateful grows proportional to some constant power of multicast group size.
- Application-driven membership has a significant impact on multicast state distribution and concentration.
- Non-branching state elimination yields up to 2 orders of magnitude reduction at even the top 10% routers or domains keeping the most local state for sparse sessions. Moreover, reduction is still substantial even at 90% session density.

Our results are applicable to future network provisioning, and useful in multicast protocol and application design.

Acknowledgments

The authors would like to thank Mark Handley for his help during the early stages of this work. We are grateful to Graham Phillips, who made available his SGB topologies and numerous graph manipulation routines. Gene Cheung, Adam Costello, Jon Crowcroft, Suchitra Raman, Sylvia Ratnasamy, Angela Schuett and Koichi Yano provided thorough feedback and valuable discussions. Many thanks to the anonymous reviewers for their insightful and constructive comments. We are indebted to them all.

References

[1] K. Almeroth. A Long-Term Analysis of Growth and Usage Patterns in the Multicast Backbone (MBone). In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[2] K. Almeroth. The Evolution of Multicast: From the MBone to Inter-Domain Multicast to Internet2 Deployment. *To Appear in IEEE Network Special Issue on Multicasting*, 2000.

[3] T. Ballardie, P. Francis, and J. Crowcroft. Core Based Trees (CBT): An Architecture for Scalable Inter-Domain Multicast Routing. In *Proceedings of SIGCOMM '93*, pages 85–95, San Francisco, CA, Sept. 1993. ACM.

[4] L. Blazevic and J.-Y. L. Boudec. Distributed Core Multicast (DCM): a multicast routing protocol for many groups with few receivers. In *Proceedings of Networked Group Communication Workshop*, Pisa, Italy, November 1999.

[5] R. Callon, P. Doolan, N. Feldman, A. Fredette, G. Swallow, and A. Viswanathan. A Framework for MPLS, Sept. 1999. Internet Draft (work in progress). Available at <http://search.ietf.org/internet-drafts/draft-ietf-mpls-framework-05.txt>.

[6] J. Chuang and M. Sirbu. Pricing Multicast Communications: A Cost-Based Approach. In *Proceedings of the INET*, 1998.

[7] S. E. Deering. *Multicast Routing in a Datagram Internetwork*. PhD thesis, Stanford University, Dec. 1991.

[8] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, and L. Wei. Protocol Independent Multicast Sparse-Mode (PIM-SM): Protocol Specification, June 1998. Internet Engineering Task Force (IETF), RFC 2362.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *Proceedings of ACM SIGCOMM*, Harvard, MA, Sept. 1999.

[10] *Georgia Tech Internet Topology Model*. <http://www.cc.gatech.edu/fac/Ellen.Zegura/graphs.html>.

[11] E. Lety and T. Turlitti. Issues in Designing a Communication Architecture for Large-Scale Virtual Environments. In *Proceedings of Networked Group Communication Workshop*, Pisa, Italy, November 1999.

[12] A. Medina, I. Matta, and J. Byers. On the Origin of Power Laws in Internet Topologies. *Computer Communication Review*, 30(2), Apr. 2000.

[13] *National Laboratory for Applied Network Research*. <http://moat.nlanr.net/Routing/rawdata/>.

[14] G. Phillips, S. Shenker, and H. Tangmunarunkit. Scaling of Multicast Trees: Comments on the Chuang-Sirbu Scaling Law. In *Proceedings of ACM SIGCOMM*, Harvard, MA, Sept. 1999.

[15] P. I. Radoslavov, R. Govindan, and D. Estrin. Exploiting the Bandwidth-Memory Tradeoff in Multicast State Aggregation. Technical report, University of Southern California/ISI, 1999. Submitted for publication.

[16] *The SCAN Project*. <http://www.isi.edu/scan/>.

[17] *The Stanford Graph Base (SGB) package*. <ftp://labrea.stanford.edu/pub/sgb/>.

[18] I. Stoica, T. E. Ng, and H. Zhang. REUNITE: A Recursive Unicast Approach to Multicast. In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[19] D. Thaler and M. Handley. On the Aggregatability of Multicast Forwarding State. In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[20] J. Tian and G. Neufeld. Forwarding State Reduction for Sparse Mode Multicast Communication. In *Proceedings of IEEE INFOCOM 98*, San Francisco, CA, March 1998.

[21] *TIERS*. <http://www.isi.edu/haldar/topogen/tiers1.0.tar.gz>.

[22] L. Wei and D. Estrin. The Trade-offs of Multicast Trees and Algorithms. In *Proceedings of the International Conference on Computer Communications and Networks (ICCCN)*, 1994.

[23] T. Wong, R. Katz, and S. McCanne. A Preference Clustering Protocol for Large-Scale Multicast Applications. In *Proceedings of Networked Group Communication Workshop*, Pisa, Italy, November 1999.

[24] T. Wong, R. Katz, and S. McCanne. An Evaluation of Preference Clustering in Large-scale Multicast Applications. In *Proceedings of IEEE INFOCOM*, Tel-Aviv, Israel, March 2000.