Building the Case for Distributed Global Multicast Monitoring

Prashant Rajvaidya
Department of Computer Science
University of California
Santa Barbara, CA 93106-5110
prash@cs.ucsb.edu

Kevin C. Almeroth
Department of Computer Science
University of California
Santa Barbara, CA 93106-5110
almeroth@cs.ucsb.edu

ABSTRACT

Key to the utility of multicast is the correct operation of the various protocols. More importantly, these protocols must operate *infrastructure-wide*. Infrastructure-wide monitoring is necessary for effectively monitoring protocol operation. Consequently, monitoring is needed for efficient management and troubleshooting of multicast networks. In this paper we present results based on the analysis and comparison of multicast routing and source announcement data collected by our global monitoring system. This system, called Mantra, collects data from various important locations in the Internet multicast topology. Our final conclusion, and not one that can be drawn by looking at any individual router, is that there exists a fair amount of inconsistency between routers in what should be consistent global state.

1. INTRODUCTION

Currently, delivery of multimedia contents over the Internet is not only being used for distribution of conventional audio and video streams, but also for services like desktop streaming, whiteboards and collaborative distance learning. To make delivery of such content a success in the Internet, several network models have been proposed, developed, and deployed. One such development is based on multicast. Multicast is a network model for one-to-many and many-to-many delivery of multimedia streams in a scalable and bandwidth-efficient manner. While a source sends data to a group, receivers explicitly join the group to receive transmitted data. To accomplish this function, multicast-capable networks have to perform special new tasks in addition to all the routing functions which conventional (unicast) networks perform. Some of these tasks include propagating information about active sources; managing distribution of data within the groups; and handling state about group participants. In the current multicast infrastructure, operation of these tasks primarily relies on three protocols[1]: the Multicast Border Gateway Protocol (MBGP),[2] the Protocol Independent Multicast (PIM) protocol[3] and the Multicast Source Discovery Protocol (MSDP).[4] MBGP acts as a route exchange protocol and allows propagation of topology information between domains. PIM uses route information to create, manage and propagate information about distribution trees. These trees consist of a set of forwarding table entries used for distribution of data from a particular source to the corresponding groups. Finally, MSDP acts as a source announcement protocol and is responsible for propagating information about active sources across the entire infrastructure.

With an increase in the amount of multimedia data, use of multicast is becoming more of a necessity than an option. Nevertheless, the extent of deployment and use of multicast has not increased significantly. One of the important factors responsible for the slow acceptance of multicast is the lack of global monitoring tools that can collect snapshots of the multicast infrastructure from multiple network domains and generate infrastructure-wide results. Such results are very important for managing multicast networks because most of the multicast protocols are inter-domain in nature and their operation requires global state information as well as inter-protocol communication. Although several useful tools exist for monitoring multicast,[5] most of them fall short of performing global monitoring and generating infrastructure-wide results. Shortcomings of the existing tools and the lack of development in this area can be largely attributed to the challenges associated with data aggregation and data processing–the two most crucial steps involved in generating a global view.[6] The challenges in this regard stem from a variety of factors including lack of support for standard data collection mechanisms like the Simple Network Management Protocol (SNMP) for multicast[7]; diversities in the data formats at different sources; frequent changes in multicast protocols; and high processing overheads.[8]

Although global monitoring is very important for effective network management in general, it becomes all the more crucial for the multicast infrastructure. While a lot of useful work has been done for global monitoring of the unicast Internet,[9,10] this is hardly the case for multicast. The role of global monitoring is more critical in the case of multicast because of several reasons. First, the visibility of sources and reachability of networks depends on how much of the global view has reached a particular network. Consequently, comparison of a local view to that the global view is important in isolating and debugging network connectivity and routing problems. Second, unlike unicast where routing problems can usually be isolated by discovering troubled links in the path, with multicast, there are potentially large numbers of receivers and a large distribution tree, spanning a wide topological area. Third, multicast distribution trees themselves are dynamic in nature. Consequently, debugging multicast routing problems requires the knowledge of global topology and global state about active sources.

In this paper we present global monitoring results based on the analysis of over two years worth of data that we have collected from several topologically important multicast routers using Mantra. While our previous work[6,8] has focussed on the challenges of collecting this data, the analysis presented in this paper focuses on using the data to support our argument that *global* monitoring is critical to a complete picture of network operation.

The rest of this paper is organized as follows. In Section 2, we introduce some important characteristics of Mantra, its data and the results we present. Section 3 presents results specific to analysis of MBGP routing tables. Section 4 presents results derived from MSDP information. The paper is concluded in Section 5.

2. RESULTS FROM GLOBAL MONITORING

Mantra is the system that we have developed to monitor the multicast infrastructure at the network layer. The results presented in this paper are based on its data. Mantra collects data at the network layer from multiple multicast routers; aggregates the collected data to generate a global view; and processes the global view to generate monitoring results. In addition, Mantra also processes data from each of the routers separately in order to generate results specific to each of the individual routers. The granularity of collected data is 15 minutes. The types of data collected includes MBGP and DVMRP routing tables, MSDP information, and forwarding state. The set of routers Mantra is currently collecting data from include: Federal IntereXchange–West (FIXW), one of the more important multicast exchange points on the West Coast of the United States; STARTAP, a core router in Abilene that acts as an interface between Internet2 and the commodity Internet; DANTE, an exchange point between the US and Dante's high speed European research backbone; and ORIX and Route-Views, routers that peer with several important US and international networks specifically for the purpose of monitoring.

The results presented in this paper focus on deployment and operation of MBGP and MSDP in the current infrastructure. The results are based on both the global view as well as various local views. Our primary focus is to (1) compare results of each of the local views, and (2) compare the local views with the aggregated global view. Through these comparisons we support our argument that local views are different. This helps to substantiate our conclusion that global monitoring is very important for the current multicast infrastructure. The aggregate global view is necessary for problem solving as well as for monitoring the global usage and deployment of multicast. Using these results, we also characterize the multicast infrastructure by analyzing the degree of network connectedness, AS path length, usage of multicast, and longer-term statistics related to routes, sources and groups.

3. MBGP RESULTS

Proper operation in the multicast infrastructure is strongly dependent on the stability of the global topology and the robustness of MBGP routing. In this section we analyze results from MBGP monitoring and derive inferences about the characteristics of the multicast infrastructure. We present two types of analyses: (1) results based on the different views of MBGP topology maps, and (2) an analysis of MBGP route statistics.

Analysis of MBGP topology maps. Topology maps act as a powerful means for gauging differences in the topology view of each router and for estimating topology characteristics like connectedness and network path lengths. The MBGP topology can be presented at a coarse (AS level) granularity and rooted at the router from where the topology data was collected. Consequently, a snapshot of the MBGP topology delineates a tree in which the end nodes are the multicast networks accessible from the router. Ideally all routers should have a view of the topology with the same set of networks and their corresponding parent ASes. However, our analysis shows that this is hardly the case and that the views available at different routers are significantly different.

Figure 1 illustrates multiple views of the MBGP topology tree as seen on March 19th 2001 at 17:30 hours (PST). The tree in the top left corner of Figure 1 illustrates an aggregate view of the MBGP topology generated on the basis of data from four different routers, FIXW, STARTAP, ORIX and DANTE. The rest of the trees depict different subsets of the AS paths that were present only in certain sets of routers. Each of the topologies is drawn only up to the level of parent ASes. Four main observations that can be drawn from these topologies are: (1) all the routers do not see the same set of AS paths, (2) the AS paths that are not present in all the topologies always terminate in a leaf AS, (3) fanout at certain ASes is very large, and (4) the maximum distance between any two nodes in the topology, i.e. the diameter of the topology, is quite small. Based on these observations we draw the following conclusions:

- *A large number of networks have poor connectivity.* Because each of the leaf ASes is a parent to several multicast networks, the absence of a leaf AS from a router's view indicates that the router does not have a path to any of the disconnected networks. Furthermore, because several leaf ASes are not present in multiple routers, a large number of multicast networks have bad connectivity to the rest of the multicast infrastructure. Consequently, sources from these networks will remain unseen by a large part of the infrastructure and the data that they are sending will likely not be accessible.
- *Varied degrees of connectivity.* The variability in the degrees of connectivity (i.e. fanout) at different ASes is very large. A detailed off-line analysis of these topologies shows that more than 62% of ASes have a fanout of 3 or less. In contrast, fanout for about 17.5% of ASes is more than 10 and a few of the ASes have extremely high fanout, in the range of 20-54. This shows that the core of the multicast infrastructure constitutes of a small number of key ASes. It can be further inferred that the multicast infrastructure is not multi-tired like the unicast and that a small number of Internet Service Providers (ISPs) and transit providers are responsible for providing multicast connectivity to a large number of end-networks. Thus, traffic is concentrated in a few places, which, in turn, can lead to a significant potential for disconnectedness.
- *Short AS paths.* Because the diameter of the topology is small, in general, the length of AS paths is very small. Consequently, all of the leaf ASes can be reached from a router with a very few AS hops. Another related conclusion is that the number of unique AS paths is also very small. Off-line analysis shows that the number of unique AS paths in the aggregate topology usually remains less than 200. Although a hop across an AS might encompass several physical hops across routers, the lack of multiple domains to cross implies fewer route flaps in the MBGP topology.

Analysis of MBGP Route Statistics. Route statistics are very useful for long-term temporal analysis and for quantitative comparison of MBGP state at different routers. Unlike topology maps (Figure 1) where results were limited to a single snapshot in time, the results in Figure 2 depict long-term statistics. Graphs in Figure 2 display results for the change in the number of MBGP routes at four different routers over time. Because each route included in the route statistics leads to a unique subnet, the number of routes seen at a router is synonymous with the number of networks that are reachable from that router. Following are the main conclusions that can be derived from these results:

- *Route variations among routers.* Route counts vary from one router to another. In addition, these variations have been persistent over the entire collection period. However, the degree of variation is not constant.
- *Route variations at each router.* The availability of routes at each router varies dramatically and frequently. This shows that from the point-of-view of all routers, multicast routing is very unstable and several networks have persistent poor connectivity.

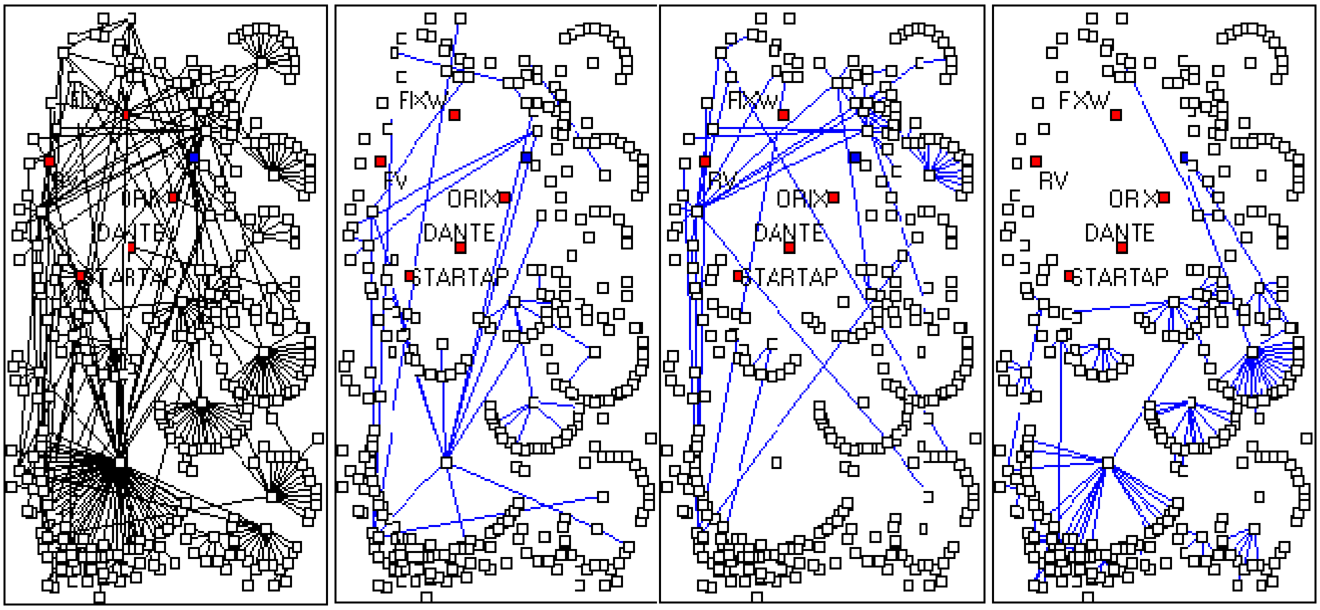

Figure 1. MBGP AS paths seen at different routers (from left to right): Aggregate; Only in ORIX; Only in STARTAP; and Only in DANTE

- *Correlation in drops.* In spite of the variation in the route counts between routers, there is still some degree of correlation. However, even this correlation is extremely inconsistent. Following are the three types of correlations that have been observed:
  - *Exact correlation.* When the routing problems are global, the number of route-drops and the time period of the drops are the same at all routers. For example, route drops across all four routers in December 2000 illustrate one such case.
  - *Partial correlation.* When the routing problems are regional, the severity of routing problems vary across routers. In such cases, route-drops occur at all the routers at approximately the same time but the degrees of the drop is not the same. One such case was observed during a two month interval between June and July 2000 when the number of routes at FIXW and STARTAP dropped and remained low. During this period, however, the number of routes at ORIX was only marginally effected.
  - *No correlation.* During some periods drops are not correlated at all and the routing problems remain local. For example, in September and October of 2000 while the number of routes at FIXW and DANTE dropped, ORIX and STARTAP remained totally unaffected.
- *Persistent, stable connectivity.* Although, the number of routes vary frequently at each of the routers and the route-drops are not always correlated, the number of routes available at a router hardly goes below 3000. This is the case because about 3000 unique networks have persistent connectivity across all the routers. This implies that there is at least a small part of the infrastructure that is consistently stable.

4. MSDP RESULTS

In the current infrastructure, MSDP is commonly used for propagating information about active sources and their corresponding groups. Although all networks should know about all global sources at any given time, this is not always the case. The set of active multicast sources varies greatly from one router to another. Therefore, having a reliable global view is very important for gauging the quality of MSDP state at any given router

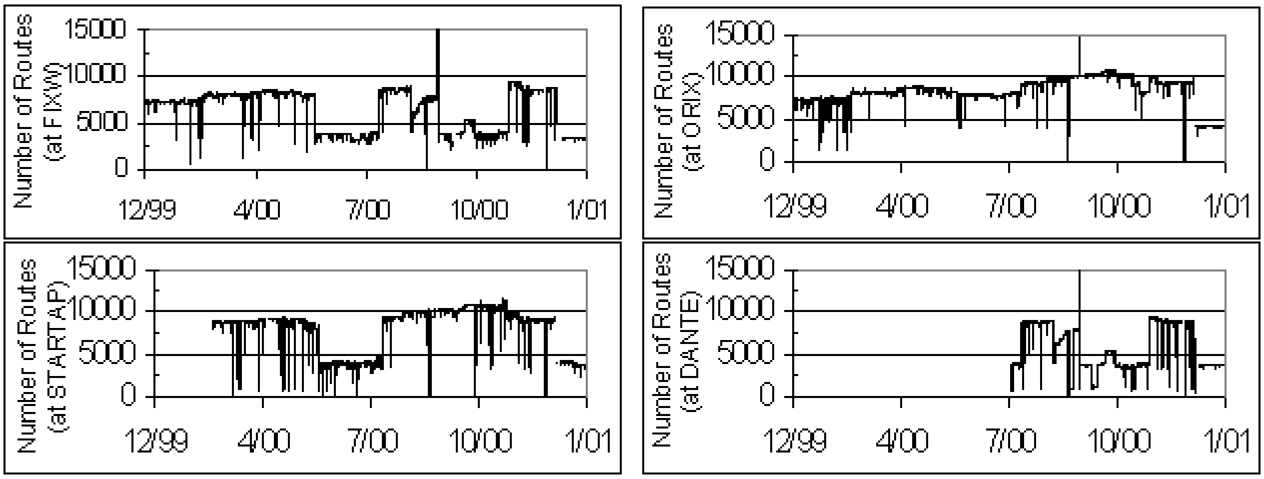

Figure 2. Number of MBGP routes as seen at different routers: FIXW(top-left); ORIX(top-right); STARTAP (bottom-left); and DANTE (bottom-right)

and, therefore, for maintaining a consistent, healthy multicast infrastructure. In addition, because the global distribution of MSDP Source Active (SA) messages is critical to building an accurate forwarding tree to all receivers, results based on the global view provide a very reliable source of information for global multicast usage statistics. Furthermore, because any inconsistencies in MSDP state is sure to cause significant discontinuities, these results are also useful for discovering and troubleshooting reachability problems. In the remainder of the section we present group statistics derived from the MSDP state that we collected. We present results based on the individual views as well as aggregate views. We use these results to analyze the differences between multiple views and to analyze the usage of multicast.

Multicast Group Statistics. Figure 3 plots the change in the number of groups announced via MSDP over time. While the top left plot shows statistics based on the aggregate view, the rest of the plots show statistics for individual routers. The following are the main conclusions that can be derived from these results.

- *Small number of multicast groups.* Except for certain occasions, the number of multicast groups available is small and stays in the range of 550-1200 groups. This shows that the usage of Internet-wide multicast is still relatively low and that the growth in usage has been slow.
- *Variation among routers.* The number of groups available at all the routers is not always the same. This shows that even the small number of groups that are active at any time are not consistently seen at all networks.
- *SA flooding.* There are several times when the number of MSDP announcements increases significantly for short intervals of time. Some of the reasons for these spikes include: (1) errors in exchange of SA state, (2) redundant SAs announced by multiple RPs, and (3) announcement of SA messages for groups that do not exist. The last reason has caused large spikes in SA state. These spikes were most severe in January 2001 when bursts of SA messages were noticed frequently (one spike reached as high as 42743 entries). All these spikes were caused by a unicast worm called the *Ramen Worm*. This worm was intending to port-scan a random set of IP addresses. However, the range of addresses included the multicast address space, and source announcements were created for each new address scanned.
- *Inconsistent correlation in changes.* Like inconsistent correlation in MBGP route drops, the changes in the number of groups seen at routers are also not strongly correlated. Like the route statistics, the level of correlation varies from *exact* to *partial* to *none*. However, as evident from the graphs in Figure 3, the instances of *exact correlations* are more frequent.

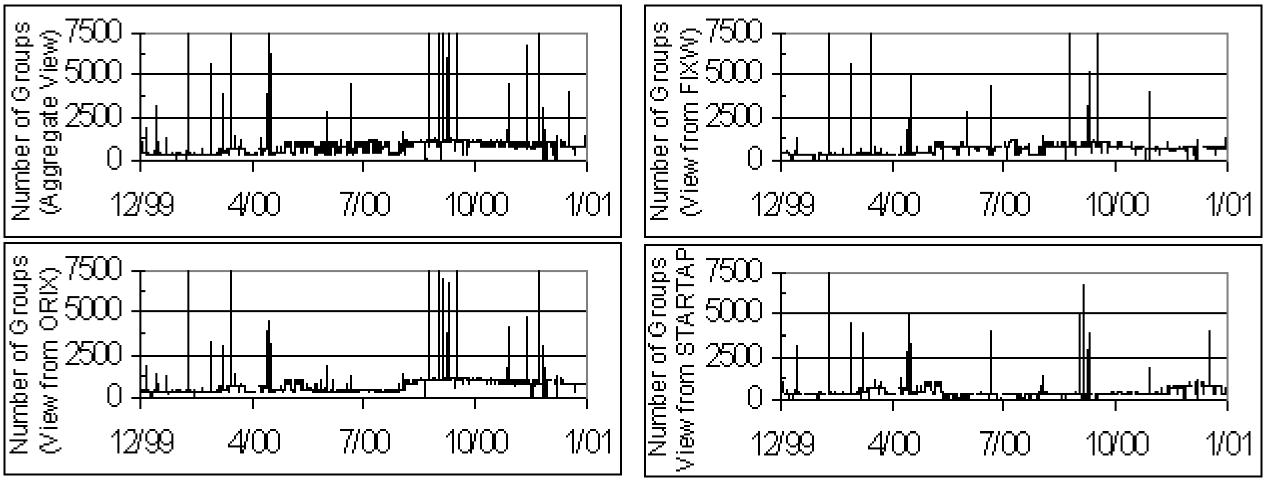

Figure 3. Number of groups announced via MSDP: Aggregate(top-left); FIXW(top-right); ORIX(bottom-left); and STARTAP (bottom-right)

5. CONCLUSIONS

Global monitoring is important for the success of multicast deployment. It provides essential infrastructure-wide data that is critical for assuring correct operation of multicast networks. The MBGP and the MSDP results obtained from Mantra further support these claims. A long term analysis of these results confirms that views of multicast networks are very diverse and that the global view of the infrastructure obtained from aggregating multiple individual views is the only reliable source for infrastructure-wide data. In addition, results from Mantra show that multicast routing is relatively inconsistent; several networks have very poor connectivity; and reachability of sources to the current infrastructure is very poor. In the future we hope to improve the robustness of the multicast infrastructure by using Mantra to isolate and fix specific problems. Furthermore, we hope to make results from Mantra useful in supporting similar efforts by multicast network engineers around the world.

REFERENCES

1. K. Almeroth, “The evolution of multicast: From the MBone to inter-domain multicast to Internet2 deployment,” *IEEE Network*, January/February 2000.
2. T. Bates, R. Chandra, D. Katz, and Y. Rekhter, “Multiprotocol extensions for BGP-4.” Internet Engineering Task Force (IETF), RFC 2283, February 1998.
3. S. Deering, D. Estrin, D. Farinacci, V. Jacobson, G. Liu, and L. Wei, “PIM architecture for wide-area multicast routing,” *IEEE/ACM Transactions on Networking*, pp. 153–162, Apr 1996.
4. D. Farinacci, Y. Rekhter, P. Lothberg, H. Kilmer, and J. Hall, “Multicast source discovery protocol (MSDP).” Internet Engineering Task Force (IETF), draft-farinacci-msdp-*.txt, June 1998.
5. K. Sarac and K. Almeroth, “Supporting multicast deployment efforts: A survey of tools for multicast monitoring,” *Journal of High Speed Networking–Special Issue on Management of Multimedia Networking*, March 2001.
6. P. Rajvaidya and K. Almeroth, “A scalable architecture for monitoring and visualizing multicast statistics,” in *IFIP/IEEE International Workshop on Distributed Systems: Operations & Management (DSOM)*, (Austin, Texas, USA), June 2000.
7. E. Al-Shaer and Y. Tang, “Toward integrating IP multicasting in internet network management protocols,” *Computer Communications–Integrating Multicast into the Internet*, pp. 473–485, March 2001.
8. P. Rajvaidya and K. Almeroth, “A router-based technique for monitoring the next-generation of internet multicast protocols,” in *International Conference on Parallel Processing (ICPP)*, (Valencia, Spain), September 2001.
9. A. Sridharan, S. Bhattacharyya, C. Diot, R. Guerin, J. Jetcheva, and N. Taft, “On the impact of aggregation on the performance of traffic aware routing,” in *International Teletraffic Congress*, (Salvador do Bahia, BRAZIL), September 2001.
10. R. Govindan and A. Reddy, “Analysis of Internet inter-domain topology and route stability,” in *IEEE Infocom*, (Kobe, Japan), April 1997.