Availability Analysis in Large Scale Multicast Networks

Falko Dressler
University of Erlangen-Nuremberg
Department of Computer Science (Distributed Systems and Operating Systems)
Martensstr. 1, 91058 Erlangen, Germany
Phone +49 9131 85-27802 / Fax +49 9131 302941
EMail: dressler@ieee.org

***Abstract* - The goal of this paper is to identify and to discuss the scalability issues of an availability analysis in large scale multicast networks as well as to introduce a solution for this task. The utilization of IP multicast in economically critical environments such as for the distribution of stock exchange prices is increasing. Therefore, the availability requirements on this service are increasing as well. Unfortunately, no measurement tools exist which determine the global multicast reachability to provide the required availability information. A possible solution is discussed in this paper, which claims to provide a high scalable measurement environment.**

***Keywords* – Parallel Algorithms and Architectures, Multicast Networks, Availability Analysis, Scalability in Distributed Systems**

I. INTRODUCTION

The increasing employment of IP multicast even for mission critical applications such as the distribution of stock exchange prices in the stock market [10] has motivated this work. To provide a high availability as required for such applications, reachability measurements have to be employed as discussed in [14, 15]. The analysis of the availability is demanded by all serious users of the global IP multicast network. Even if IP multicast helps to save resources for a one-to-many transmission over the internet [11], there are a few drawbacks hindering multicast to get employed for mission critical applications. One problem in the current global multicast infrastructure is the absence of adequate measurement tools.

Even if there are first approaches to test the functionality of the multicast network since the early beginnings of the development of IP multicast routing protocols, all these concepts do not allow a scalable deployment over large parts of the network or even in the global multicast enabled internet [2].

The typical problems of testing the availability of multicast networks [5] are introduced in this paper accompanied by a short overview to the state of the art in availability measurements in multicast environments. Additionally, a new approach is presented, named multicast quality monitor (MQM). The concepts of this idea allow a high scalable availability analysis even in large scale multicast networks [7]. These concepts are discussed in the main part of this work. A conclusion summarizes this paper including some outlook to other ongoing work.

A. Scalability Issues

Scalability is always an issue in multicast environments due to the working principles of multicast routing [16, 17]. The basic concepts of multicast are as follows: the sender of a packet stream sends its packets only once to a so-called multicast address and the network is responsible to deliver the message to each client who is interested in receiving traffic for this particular multicast address. Therefore, each of these clients receives a copy of each packet sent to the multicast group. The scalability of an availability test strongly depends on the concept of the message passing between all test-stations that will be called probes in the following. The easiest approach is to have each participating probe sending test packets on a regularly basis. These test packets might be responded by sending answer packets back to the originator or – which allows more precise measurements – to the multicast group allowing each other probe to receive this response.

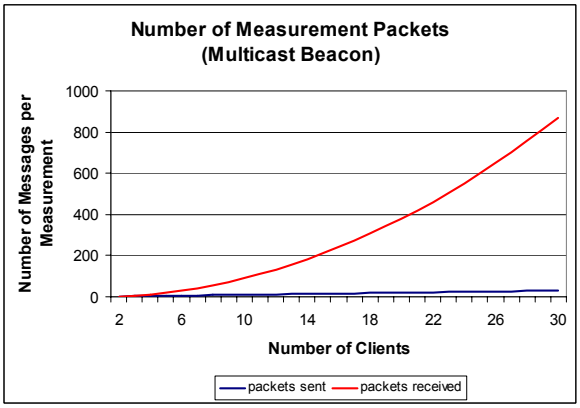

Fig 1. Scalability of a reachability test using the multicast beacon. Shown is the relationship between an increasing number of clients and the resulting number of messages per measurement.

An example for such a measurement is the multicast beacon which is described later in the related work section. The scalability of this approach is shown in Fig 1. It can be seen that the number of messages that have to be received and analyzed at each participating client increases

dramatically according to the formula $n = p * (p - 1)$ where p is the number of involved probes. The amount of wasted network utilization b can be calculated as follows:

b_{stream} = packet rate * packet size

$b_{total} = b_{stream} * n$.

For example, the multicast beacon sends about 10 packets per second with a size of 100 bytes each. Thus, the resulting bandwidth requirements b_{total} at each probe for 30 participating clients is about 7 Mbit/s which is too much just for measurement traffic besides the regular network usage.

B. Completeness of Availability Analysis

Besides the scalability, there is an issue of importance depending on the measurement concept as well: the completeness of the analysis. In case of failures and partially unavailable network parts, i.e. split sub-networks, the knowledge about the internal behavior of each of these parts is required in order to provide a complete analysis.

Due to the high complexity of multicast routing protocols and due to incomplete and non-interoperable implementations, network partitioning appears to be very common in well-known multicast networks [13]. The problem is described in more detail in Fig 2.

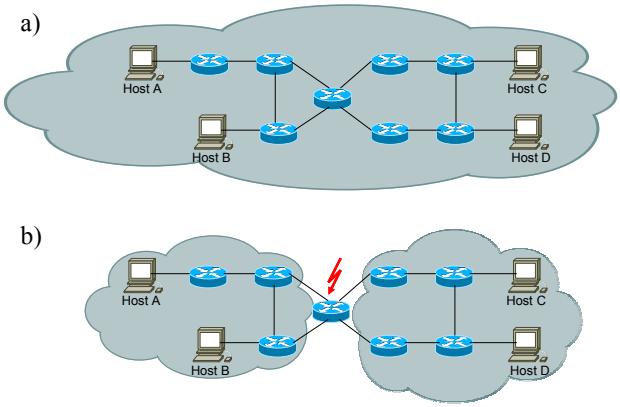

Fig 2. Network partitioning. The functional multicast network (a) can be split into several partitions (b) due to a failure in multicast forwarding at some point in the network. It is up to the measurement system to analyze this behavior and to provide information about each particular sub-network.

Typically, it is not possible to get information about a (sub) network if there is no measurement probe deployed in it that is actually sending test packets. Therefore, most approaches are based on the concept of building probes that periodically send data packets and reply on received requests. Thus, the scalability issues discussed in advance apply again. New concepts have to be developed which allow a high scalability accompanied by a complete availability analysis. One possible solution is discussed in section III.

II. Related Work

The discussion of the related work must be started with the already aged tool mtrace. It has been developed in the early beginnings of the MBone [8] in order to test the functionality. Additionally, the multicast reachability monitor introduced by Almeroth [3] and the multicast beacon [4] from the NLANR (National Laboratory for Applied Network research) are investigated.

A. mtrace

Mtrace is one of the most well-known tools to test the reachability including a full trace of the used multicast path between two systems. A special feature has been built into the multicast routers [9] to make mtrace work. Despite the fact that mtrace is a very useful tool to check the multicast connectivity and routing, it also has its shortcomings. Therefore, it may happen that mtrace does not trace the path successfully even if the multicast forwarding is working properly, just because there are routers on the path, which have not implemented the mtrace features or at which the feature is administratively disabled. Therefore, this tool can be used for single tests only and, typically, only in very small and clearly arranged networks.

B. Multicast Reachability Monitor

The multicast reachability monitor (MRM), formerly known as the multicast route monitor was developed to allow a centralized reachability measurement based on probes located all over the multicast network. End systems can be used as probes as well as the multicast routers themselves. The MRM, which started as an IETF draft [1], defines three different processes: the MRM manager, the test sender, and the test receiver. Controlled by the manager, the multicast reachability monitor is able to create a configurable packet flow at each test sender. Using the received packets, the test receivers are able to compute measurement results, such as the packet loss ratio, which provide a good estimation of the reliability of the network.

Comparing the abilities of the multicast reachability monitor with the described problems or more precisely the scalability and the completeness it has been shown that the deployment of the MRM can be either in some degree scalable, i.e. if only one or a few senders are implemented. Unfortunately, network partitioning might not be recognized in this case and the connectivity can only be tested in one way from the sender towards to the receivers. On the other hand, the MRM has strong scalability problems if all the receivers are working as senders as well. In this case the network utilization for the measurement is much too high.

C. Multicast Beacon

The multicast beacon is the result of a research project from the NLANR. Currently, there is an implementation in JAVA for the so-called beacon clients available, which should run on nearly every end system with an installed JVM (java virtual machine). The so-called beacon server consists of a perl program. The principles of the multicast beacon and the MRM are very similar. The definition of the multicast beacon includes a server computing the QoS parameters from

measurement results and the clients, named beacons, which are sending and receiving the measurement packets. All the beacons interact directly with each other by constantly sending IP multicast packets to an administratively configured multicast group. Each beacon client reports its measured data, i.e. the results of received packets (beacons) to the server. The server calculates a matrix including each active client and allows these results to be accessed via a web gateway.

The main differences between the MRM and the multicast beacon are the capability of the multicast beacon to provide a direct access to the measurement results and the wider range of QoS measurements (packet loss ratio, delay, and jitter). On the other hand, the MRM allows one to distinguish between a test sender and a test receiver. This differentiation results in a much lower impact on the network, especially if broadcasting scenarios are the most common applications in the particular network under study. The scalability was already discussed during the introduction. It was shown that this approach is not scalable for application in very large scale multicast networks.

D. Summary

As shown above, none of the available tools is able to fulfill the requirements for a complete and scalable availability analysis. New concepts are required. An approach to solve the problems is described in the following section.

III. A New Approach: Multicast Quality Monitor

The focus of this section is a new multicast ping mechanism introduced as part of the multicast quality monitor [6]. The primary goal of this new methodology is a high scalability. The MQM ping mechanism was designed for a complete analysis of the reachability and, therefore, of the availability of large scale multicast networks.

A. MQM Ping Mechanism

The MQM ping mechanism is directly based on the working principles of IP multicast.

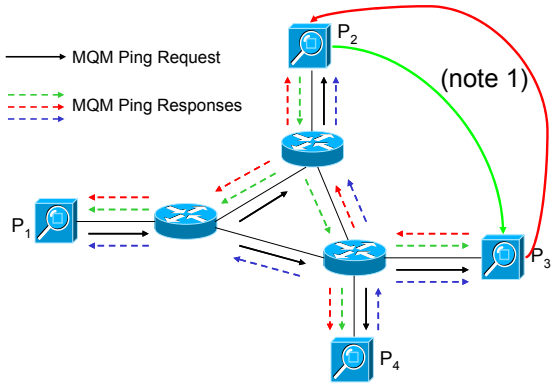

Fig 3. MQM ping mechanism. Without restricting the generality, we believe that probe P_1 is sending a MQM ping request packet. Based on the working principles of IP multicast, all the other probes receive this request and start sending a response message. Therefore, P_1 gets an answer from each participating probe and is able to analyze the reachability in the network. As shown in note 1, e.g. P_2 receives an answer from P_3 as well as vice versa. Using these receives response messages, P_2 and P_3 can analyze the behavior of the network between them. Thus, using a single ping request it is possible to analyze the complete multicast network.

As shown in Fig 3 the MQM ping relies on the transmission of ping request messages and the reaction to the reception of such packets, the transmission of corresponding response messages.

In contrast to other tools such as the mentioned multicast reachability monitor or the multicast beacon, the MQM ping is not only based on a request-reply mechanism. It allows to use replies only to get enough information about the connectivity in the network to provide a complete analysis of the availability.

The goal of a high scalability accompanied by a high fault tolerance is achieved by using the following concepts:

- there must be one MQM ping request per minute (all the typical multicast routing protocols have a timeout for the entries in their routing tables of 3 minutes, thus, the states must be refreshed within this interval)
- there must be two MQM ping requests per (sub) network (the messages are unreliable and, therefore, may get lost)

This can be achieved in an implementation using the following methodology:

- at the startup, a single MQM ping request is sent (enabling all the other probes to learn about the new participant)
- in the following a MQM ping request is sent only if there were less than two requests received in the last interval (in the last minute)

Regarding the scalability of the MQM ping mechanism it can be said that a proportional scaling can be achieved. A graph showing the scalability is provided in Fig 4.

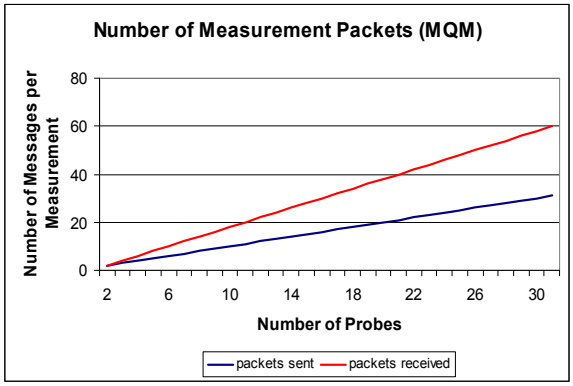

Fig 4. Scalability of the MQM ping mechanism. Obviously, the increase of required messages is no longer exponential with an increasing number of probes as shown for the multicast beacon. Now we achieved a proportional scaling which is applicable even for large scale networks.

The novelty of the shown ping mechanism is its ability to work in a multicast environment without the common problem of packet explosion. Based on the working principles of the new multicast ping it is possible to detect network partitionings (in each partition must be ping requests sent by any participating probe). Additionally, there is the requirement to sent further requests if some request packets got lost due to overload situations.

In summary it can be said that the MQM ping mechanism allows a high scalable reachability measurement of IP multicast networks, i.e. it reduces the impact on the network to a very low level. For example, if 30 probes are employed, a theoretical peak in network utilization due to the overhead of the measurement of about 46 kBit/s is feasible (100 Byte per packet, 2 requests and 29 responses per minute). In practice, not all packets in each interval will be sent at the same second and, therefore, the wastage of network resources will be considerably lower.

B. Availability Analysis

Normally, the reachability should be maintained by the simplest possible mechanisms. Redundancy is provided in typical backbone networks. Even if the provisioning of the connectivity sounds easy, especially in IP multicast environments this cannot be presumed. The multicast reachability suffers from the complexity of the multicast routing protocols and the lack of experience of network administrators declining with these mechanisms. Another problem is the still miserable interoperability between devices of different manufacturers and, partially, the incomplete implementation of the protocol stacks.

Using the results of reachability measurements over a period of time, the reliability of the network can be calculated. High availability systems require a reliability of nearly 100%. Therefore, reachability means connectivity at a certain point of time and reliability stands for the percentage reachability over a period.

Based on the reachability measurements using the MQM it is possible to estimate the availability of single network paths as well as of large network parts. The results of the single measurements are distributed over all the employed measurement probes. This results in a typical problem in distributed systems, how to retrieve the required information with a minimum transmission overhead. This kind of problem was discussed in many facets, particularly in peer-to-peer networks. We also thought on using IP multicast for the collection of the measurement results but decided to use reliable unicast transmissions based on results of [12] and due to the required overhead for performing the reliable transmission.

Depending on the amount of measurement data, it seems to be advisable to preprocess the information at each probe before they are transmitted to a common analyzer. Finally, the information must be centrally collected and analyzed. For example, it is possible to conclude the availability of a network path by looking for received messages along the path. It must be considered that a received packet can only be used to determine the unidirectional connectivity between the two associated probes.

An example of such an analysis is shown in Fig 5. The connectivity measured by the probes P_1-P_4 is noncontiguous as shown by the crossed unidirectional arrows. The conclusion of investigating the figure is that there must be a problem in multicast forwarding between the routers R_3 and R_4 or, maybe, only at R_4. If there is a recording of the network behavior available over a period, information can be gathered whether there is a temporary failure or not.

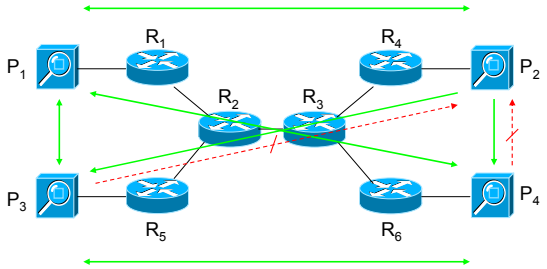

Fig 5. Availability Analysis. In this sample network, there were no correctly received measurement data packets between P_4 and P_2 as well as between P_3 and P_2. The conclusion is that there must be a failure between R_3 and R_4 or, maybe, only at R_4.

C. Summary

In summary it can be said that the analysis of the availability of an IP multicast network using the mechanisms of the multicast quality monitor is highly scalable and fault tolerant. Thus, there are mechanisms available that allow an examination of multicast networks to determine the reachability for example to verify a SLA (service level agreement) which guarantees some degree of availability for a particular network connection or even for a larger part of the network.

IV. Conclusion

In this paper the requirements for availability examinations in multicast networks were shown. Especially new services such as the transfer of stock exchange prices and other mission critical applications demand a high availability of the employed network. During the discussion of the related work, it was shown that none of the former approaches offers a complete and scalable solution for such measurements.

The multicast quality monitor includes a new methodology that allows a high scalable reachability analysis which covers all the parts of the possibly split network. The MQM ping mechanism was closely analyzed and is was shown that it allows a statement on the availability of single network connections as well as of larger network parts. The high scalability of this approach allows the involvement of a large number of measurement probes without an exorbitant impact on the network.

The final question is “Quo vadis?” The collection of the measurement results can be further optimized. New

approaches known from peer-to-peer networks can be employed for this task. Nevertheless, it must be said that the most important problem, the measurement itself, can be assumed to be solved by the new methodology, the MQM ping mechanism.

REFERENCES

[1] K. Almeroth, L. Wei, and D. Farinacci, "Multicast Reachability Monitor (MRM)," draft-ietf-mboned-mrm-01.txt, July 2000.

[2] K. C. Almeroth, "The evolution of multicast: from the MBone to inter-domain multicast to internet2 deployment," IEEE Network Magazine, vol. 14, January/February 2000.

[3] K. C. Almeroth, K. Sarac, and L. Wei, "Supporting Multicast management Using the Multicast Reachability Monitor (MRM) Protocol," UCSB, Technical Report TR 2000-26, May 2000.

[4] K. Chen, M. Kutzko, and T. Rimovsky, "Multicast Beacon Server v0.8.X (Perl)," January 2002. (http://dast.nlanr.net/projects/beacon)

[5] F. Dressler, "How to Measure Reliability and Quality of IP Multicast Services?," Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (IEEE PACRIM'01), vol. 2, Victoria, B.C., Canada, August 2001, pp. 401-404.

[6] F. Dressler, "An Approach for QoS Measurements in IP Multicast Networks, MQM - Multicast Quality Monitor," Proceedings of International Network Conference (INC 2002), Plymouth, UK, July 2002.

[7] F. Dressler, "MQM - Multicast Quality Monitor," Proceedings of 10th International Conference on Telecommunication Systems, Modeling and Analysis (ICSTM10), vol. 2, Monterey, CA, USA, October 2002, pp. 671-678.

[8] H. Eriksson, "MBONE: the multicast backbone," Communications of the ACM, vol. 37, pp. 54-60, August 1994.

[9] B. Fenner and S. Casner, "A "traceroute" facility for IP Multicast," draft-ietf-idmr-traceroute-ipm-07.txt, July 2000.

[10] N. F. Maxemchuk and D. H. Shur, "An Internet multicast system for the stock market," ACM Transactions on Computer Systems (TOCS), vol. 19, pp. 384-412, August 2001.

[11] P. V. Mieghem, G. Hooghiemstra, and R. v. d. Hofstad, "On the efficiency of multicast," IEEE/ACM Transactions on Networking (TON), vol. 9, pp. 719-732, Dec. 2001.

[12] T. Oh-ishi, K. Sakai, K. Kikuma, and A. Kurokawa, "Study of the Relationship between Peer-to-Peer Systems and IP Multicasting," IEEE Communications Magazine, vol. 41, pp. 80-84, January 2003.

[13] M. Ramalho, "Intra- and Inter- Domain Multicast Routing Protocols: A Survey and Taxonomy," IEEE Communications Surveys & Tutorials, vol. 13, 2000.

[14] K. Sarac and K. C. Almeroth, "Monitoring reachability in the global multicast infrastructure," Proceedings of International Conference on Network Protocols (ICNP), Osaka, Japan, 2000.

[15] K. Sarac and K. C. Almeroth, "Supporting the Need for Inter-Domain Multicast Reachability," Proceedings of Network and Operating Systems Support for Digital Audio and Video (NOSSDAV '00), Chapel Hill, NC, USA, June 2000.

[16] A. Striegel and G. Manimaran, "A Survey of QoS Multicasting Issues," IEEE Communications Magazine, vol. 40, pp. 82-87, June 2002.

[17] T. Wong and R. Katz, "An Analysis of Multicast Forwarding State Scalability," Proceedings of 2000 International Conference on Network Protocols, 2000.