AN OVERLAY SOLUTION TO IP-MULTICAST ADDRESS COLLISION PREVENTION

Piyush Harsh, Richard Newman
Computer and Information Science and Engineering
University of Florida, Gainesville FL 32611
U.S.A.
{pharsh, nemo}@cise.ufl.edu

ABSTRACT
Multicast applications before they start transmitting content must choose a multicast channel on which to transmit. Unlike IP unicast addresses, multicast addresses are not long lived entity. Moreover many applications can choose to transmit data on the same channel. Every application which chooses to transmit data intended for receivers outside its own administrative domain must choose a globally scoped channel. Since most of globally scoped multicast channel addresses are not statically assigned, there is a high probability of address collision among applications if they are assigned addresses randomly and without the prior knowledge of global assignments of such addresses. This paper proposes an overlay solution to dynamic "globally scoped" multicast address allocation and collision prevention.

KEY WORDS
Multicast, Scoping, Overlay, Address Allocation, Collision Prevention.

1. Introduction

With the improvement in wire speed and bandwidth capacity, the Internet is seeing a large proliferation of rich multimedia content. Many of these streams adhere to classroom style lecture model of content delivery where there is one content source and multiple receivers. IP Multicast is very well suited to support this kind of content delivery model. IP Multicast also significantly reduces bandwidth wastage at the cost of increased complexity in the routers. Also it provides seamless scalability in terms of subscribed user-base. This provides a huge incentive to the multimedia content provider who no longer has to upgrade its servers and allocate more bandwidth with increasing user base. With ASM model of IP Multicast still many years away from being phased out and IGMP v 3[1] still in drafts stages, there is a need for a multicast address allocation scheme which can scale globally and yet provides reasonable guarantees against multicast address collisions.

Within IPv4, multicast addresses have been traditionally designated as class D addresses. These addresses range from 224.0.0.0 through 239.255.255.255. For the purpose of packet forwarding these addresses does not have any mask associated with them. In this paper CIDRized notations are used just for the purpose of denoting a group of address ranges. Of the above address range, not all addresses are available for usage by the content providers. Some of these addresses have been reserved for internetwork control. IANA is central authority that maintains strict control on how these addresses are used. Of the various multicast address ranges assigned for numerous purposes by IANA, only 224.2.0.0/16 which is the SAP[2] / SDP[3] range, 232.0.0.0/8 which is reserved for SSM[4] sources and 233.0.0.0/8 which is AS-Encoded, statically assigned GLOP[5] range can be used to transmit multicast traffic on global scope.

If a host just wants to transmit a stream within a particular administrative domain, then it can choose an address from administratively scoped multicast address range which is 239.0.0.0/8. Allocation within this range can be handled very easily using the same protocol as in DHCP[6]. Problem may arise if a host wants to multicast a traffic stream all over the Internet crossing its own administrative boundaries. The host application must choose a globally scoped multicast channel address in such a way so as to minimize the likelihood of that address already being used elsewhere in the Internet. If such precautions are not taken, it may result in huge amount of cross-talk and could result in service deterioration.

The paper in subsequent sections explains some of the existing proposals that try to address the globally scoped address collision prevention problem along with the benefits and drawbacks of each scheme. That is followed by the description of the hierarchical structure proposed by the author. A brief worst case delay analysis of our address allocation algorithm is presented, followed by anticipated benefits and drawback of the proposal. The paper ends with a brief discussion on future research direction as well as current work in progress by the authors.

2. Design Goals

We did some analysis on the desired qualities that any global service architecture proposal must incorporate. We now list some of those in decreasing order of their relevance in the currently deployed infrastructure context.

Deployment on the existence infrastructure

Any new global service architecture today must be deployable on the existing internet architecture. Regional ISPs and core-network service providers have already spend billions of dollars on currently installed hardware and any proposal that requires large scale hardware upgrade will most likely not be implemented by the service provider today.

Scalability

Any IP Multicast address allocator architecture must be globally scalable. The allocator service can be used by many users all over the world spread throughout the Internet.

High Availability

Address allocator architecture must be designed keeping intermittent link/equipment failures in mind. Failures in any link or hardware in a particular sub network must not bring the whole Internet to its knees. Service in other networks should not be affected by the partial failure in other domains.

Resilience against DDoS

With the increasing number of global occurrences of DDoS attacks on highly popular internet services, protection against DDoS attacks against such services is becoming more and more desired.

Low Bandwidth Usage

Internet services like address allocator services or session announcement services that facilitates in the proper functioning of other consumer services must not consume significant portions of bandwidth for its internal control messages. The bandwidth utilizations should be low so as to not affect the other services / communications.

3. Analysis of existing tools and other proposed solutions

MBONE tool *sdr* is still in use by some applications for address allocation for a newly created multicast session. For a globally scoped session, *sdr* allocates address randomly selected from the SAP/SDP[3] range 224.2.0.0/16. While random allocation scheme is simple and easy to implement, it does not scale well as number of sessions increase. There are bound to be address clashes in truly random allocation schemes.

'sdr' alleviates some of the allocation woes by using informed random multicast allocation or IRMA. This introduces an additional problem of global session state information which must be maintained by the sdr tool. This scheme might work for small number of sessions in a smaller multicast scope. And the effectiveness of such a scheme is heavily dependent on the session announcement message delays and packet loss rates on the Internet. And on the global scale, maintaining individual session states is truly impractical.

IPRMA or Informed Partitioned random Multicast Address Allocation scheme which was proposed by Van Jacobson [7] was a partial improvement on reducing address collision while allocating session addresses locally. In [8] the author shows that depending on the number of partitions in IPRMA, the address collision rate varied in between $O(\sqrt{n})$ and O (n) where 'n' is the number of addresses available for allocation. The optimal rate of O (n) was achieved in the case where no two TTL values fell in the same partition. Ideally this would suggest having as many partitions as there could be different TTL scopes for various multicast sessions. This introduced effective utilization problems where one of the higher demand partitions would become full while other partitions remaining underutilized.

MASC / BGMP architecture for hierarchical and dynamic multicast address allocation has been proposed in [9]. MASC proposal has lots of nice features such as global scalability. Its hierarchical address prefix allocation scheme gels well with CIDRized philosophy on network address assignments. Their scheme also results in compact routing table and less third party dependence for efficient multicast routing. One nice feature is the multicast tree being rooted in the domain owning the multicast prefix chunk.

MASC[9] protocol wait period of almost 48 hours before claiming a set of addresses could result in potential collision related instability on the global scale. Also threshold based address claim mechanism seems defensive algorithm at best. Because of 48 hour wait period before claiming a address set, there could be instances where in MASC the MAAS servers must resort of random address allocation to requesting sessions even though there might still be available free addresses in the parent's address set.

In [10] the authors presented a very comprehensive analysis of the multicast address allocation problem. They compared simulation results of various allocation algorithms including MASC[9], Cyclic [11] and MaxQ [10] and found that surprisingly prefix based allocation

schemes did equally well compared to contiguous allocation schemes. Their simulation also pointed that allowing just 2 address chunks to be owned by sub domains in MASC protocol was too restrictive and in fact with 4 chunks allowed the overall allocation performance to improve significantly.

4. Hybrid Overlay-Multicast Address Allocator (HOMA)

Our proposed multicast address allocator scheme tries to overcome some of the shortcomings of MASC[9] proposal by incorporating some recommendations by researchers in [10] and making use of a hybrid hierarchical overlay network of address allocator servers on the similar lines of MASC proposal. In addition we augment the architecture with sub-domain level node peerage using dedicated multicast channels at each level in the hierarchy. We conjecture that our proposed architectural modification should result in better address space utilization while trying to minimize routing flux at the global level at the cost of slightly higher routing flux at the lower level routers. Our proposal also tries to retain the global address allocation on the lines of unicast CIDRized scheme as much as possible on the similar lines of MASC. But we try to improve on the latency by forgoing claim-collide scheme for request-reply model.

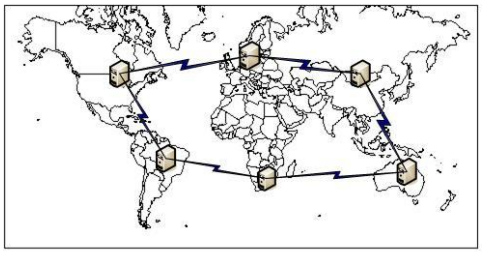

Figure 1: Global TLDs Overlay

In our design, IANA initially assigns the globally scoped multicast addresses among global TLDs. This division might take into consideration global statistics on multicast session's usage pattern and address demand. IANA involvement in our scheme is only limited to this initial address allocation to each of the TLD.

Each global TLD serves as the root level domain for the regional and enterprise domains under its jurisdiction.

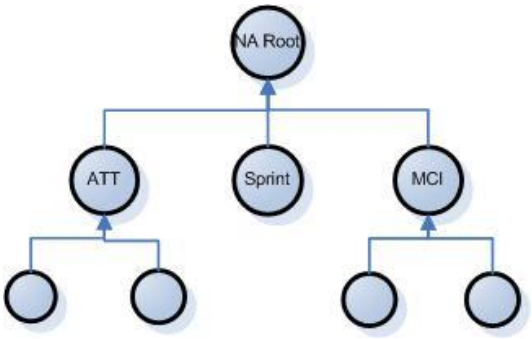

Figure 2: ISP Tree rooted at global TLD

In order to maximally utilize the multicast addresses, each sibling domain at any level also forms a dedicated peer network which could be an IP overlay or using a multicast channel. Necessary information for form the overlay peerage could be transmitted to each of the siblings at the next layer by the parent node. For instance in the example tree hierarchy above, ATT, Sprint and MCI forms a peer network among themselves. This peerage network is constructed at each level among the sibling nodes at that level in the tree hierarchy.

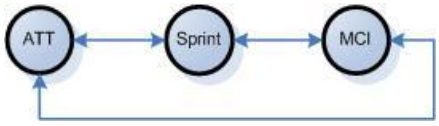

Figure 3: Peer n/w among sibling nodes

5. HOMA Address Allocation Algorithm

Each of the nodes in HOMA framework maintains two parameters α and β from the time an address block is allocated to it from the parent node until the time when the block lease expires. The values of α and β are updated every 5 minutes duration.

Let λ be the number of new address requests within current 5 minute time slice. Let μ be the number of address release by multicast applications within the same time frame. Then –

$$\alpha_{new} = \lambda.p + \alpha_{old}(1 - p)$$
$$\beta_{new} = \mu.p' + \beta_{old}(1 - p')$$

where parameters p and p' are experimentally determined. The parameters α and β are used as an estimate of future rate of new address requests and release of old addresses respectively.

Also let γ denote the address utilization factor at each node that when reaches the predetermined threshold value, would trigger the additional address request protocol within the HOMA node.

The additional address requirement can be computed as follows –

$$\text{Let } N = [\text{lease time} - \text{current time}] \div 5$$

Here N represents the number of 5 minute slots until the current address set allotted to this HOMA node expires.

Then addition addresses anticipated δ is given by

$$\delta = [(\alpha - \beta) \times N] - \#\text{free\_addresses\_remaining}$$

Let us assume that first time a HOMA node is brought online it directly contacts the parent node for a chunk of multicast address, it gets the sibling peerage details from the parent node and joins the sibling peer network. All this can be considered part of the HOMA bootstrapping process.

Pseudo-code for address allocator module –

If incoming request is for a new channel address by a multicast application –

- If a free channel address is available then allocate the address to the requesting application after negotiating the address lease time properly.
    - Update γ, λ
- If a free channel address is not available, then allocate a channel address randomly from the parent's address space.
    - Update λ

If incoming request is to release one of the already allotted addresses by a multicast application –

- If the address belongs to the set owned by this HOMA node, then add it to the free address list.
    - Update γ, μ
- If the address does not belong to the address set owned by the HOMA node, do not add to free address list
    - Update μ

At every 5 minutes interval –

- Recompute α, β
- Set $\lambda = \mu = 0$

After every address allocation / de-allocation check the value of updated γ.

- If $\gamma <$ threshold: Do nothing.
- If $\gamma \geq$ threshold
    - Compute the anticipated additional address required δ
    - If $\delta > 0$, initiate a request for δ number of addresses on the sibling peer network and wait for 2 minutes for responses.
        - If any response comes, add addresses to the free address pool keeping track of the lease associated with those addresses.
        - If no response comes, initiate additional address request to parent HOMA node.

If additional address request is received on the sibling peer network –

- Compute possible disposable address count ϕ using the following relation:

$$\Phi = \#\text{free\_addresses\_remaining} - [(\alpha - \beta) \times N]$$

    - If $\phi > 0$, indicate willingness to allocate ϕ set of addresses to the sibling node. Treat this allocation just like any other address allocation.
    - If $\phi \leq 0$, then do nothing.

This pseudo-code is implemented at each HOMA node and each node executes this pseudo-code independently of one another. There is no centralized component in the above pseudo-code.

6. Time-delay Analysis of HOMA algorithm

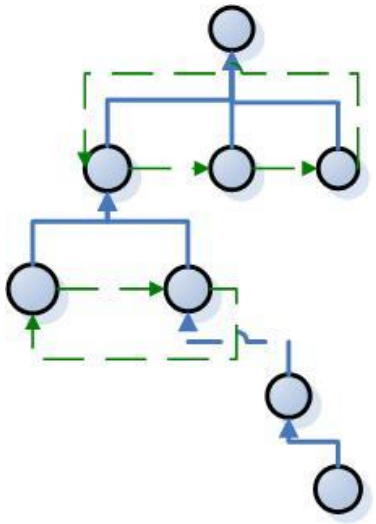

Figure 4: A general scheme of HOMA nodes

For purpose of doing time delay analysis suppose that that with probability π the additional address demand is satisfied from one or more sibling nodes. In the worse case any node must wait for a duration of 2 minutes before sending additional address request to it's parent node, we can define a recursive equation for the overall delay in terms of tree depth'd'.

$$\text{Delay} = 2\pi + (2 + \Lambda_d)(1 - \pi)$$

where Λ_d is the delay if the request must be made to one's parent node.

$$\Lambda_d = 2\pi + (2 + \Lambda_{d-1})(1 - \pi)$$

Here in the above equation, Λ must also account for time delay in locating a possible chunk of address in ones internal free addresses list.

The value of π remains to be experimentally determined. It can be calculated by tracking the fraction of cases during any simulation run where the additional address demand was satisfied by sibling nodes. We conjecture that this delay behavior is more suitable for a dynamic session scenario than delay behavior of claim-collide mechanism in MASC proposal. Whether this is valid remains to be seen.

7. Advantages of HOMA Architecture

Since HOMA distributed algorithm can be implemented in software, there is no need for ISPs to update their routing hardware. Ability to exist in current deployed environment is one of the greatest strengths of proposed algorithm. Lack of any centralized components in the proposed algorithm is in line with accepted trend in the Internet management protocol design community. It also makes the algorithm robust against localized failures.

The fact that the global TLDs are well known in our design; it could be used effectively to design simple DDoS prevention strategy. The child nodes of first few level deep parent nodes could also be assumed well known thereby enabling the parent node to filter out protocol messages from downstream clients more effectively. Resilience against DDoS attacks in the Internet management architecture is becoming more and more important.

Another important feature in HOMA design is minimization of routing flux. Since the address allocation is hierarchical in our proposal, any address sub lease and exchanges among the sibling nodes would result in routing table entry changes in the sibling HOMA nodes' parent node and no higher. Routing stability is paramount for any globally scalable Internet service architecture.

Since the address allocation scheme of HOMA proposal is more responsive and real time by design, there is no need for using a very conservative threshold setting as proposed in MASC paper. We conjecture that using non conservative threshold in HOMA algorithm would result in better address space utilization. One of the reasons for this is improved address reuse among siblings and HOMA being open to both chunks as well as individual address allocations to child / sibling nodes. How this relaxation translates to architectural complexity compared to other proposals still remains to be studied.

8. Current and Near Future Work

Our next goal is to implement the simulation of HOMA algorithm and do comparative analysis with MASC and other proposals. We are currently in the process of implementing a valid simulation model for our design. We are also trying to come up with a theoretical analysis with respect to protocol stability of HOMA design using stochastic techniques such as birth and death processes [12].

As stated earlier that algorithm parameters such as p, p' and π are to be experimentally determined and those remain to be our prime candidate to be determined during the simulation phase. It would be somewhat interesting to find the stability range for these parameters with respect to HOMA design.

Although HOMA design has been proposed for dynamic Multicast address allocation, its use is not limited to just Multicast. We believe the general idea in principle could be applied to many other domains where limited addressing could be a challenge and possible bottleneck.

Acknowledgements

The authors would like to extend their gratitude towards Dr. Randy Chow of CISE Department who took some time out of his busy schedule to review the early drafts of this paper. His comments were very helpful in refinement of the HOMA architecture in its early stages.

We are also thankful to fellow graduate students in Center for Operating Systems, Networks and Security at CISE, University of Florida who participated in many discussions and presentations on HOMA. We would especially like to thank InKwan Yoo and Panoat Chuchaisri for their critical analysis and suggestions.

References

[1] H. Holbrook, B. Cain, B. Haberman - "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast". IETF RFC-4604, August 2006.

[2] M. Handley, C. Perkins, E. Whelan – "Session Announcement Protocol", IETF RFC-2974, October 2000.

[3] M. Handley, V. Jacobson – "SDP: Session Description Protocol", IETF RFC-2327, April 1998.

[4] S. Bhattacharyya – "An Overview of Source Specific Multicast", IETF RFC-3569, July 2003.

[5] D. Meyer, P. Lothberg – "GLOP Addressing in 233/8", IETF RFC-3180, September 2001.

[6] R. Droms – “Dynamic Host Configuration Protocol”, IETF RFC-2131, March 1997.

[7] Van Jacobson – “Multimedia Conferencing on the Internet”, *SIGCOMM* ‘94

[8] Mark Handley – “Session Directories and Scalable Multicast Address Allocation”, *SIGCOMM* ‘98

[9] Satish Kumar, Pavlin Rodoslavov et al. – “The MASC/BGMP Architecture for Inter-domain Multicast Routing”, *SIGCOMM* ’98.

[10] Daniel Zappala, et al. – “Special Issue of Computer Networks”, *Elsevier Science* ‘04

[11] Marilynn Livingston et al. – “Cyclic Block Allocation: A New Scheme for Hierarchical Multicast Address Allocation”, *Networked Group Communication*, 1999, 216-234.

[12] Sheldon M. Ross – “*Introduction to Probability Models*”, 4th Edition, 251.