Internet Engineering Task Force Jon Crowcroft

INTERNET DRAFT
draft-crowcroft-mat-00.txt
November, 2001
Expires: April, 2002

Multicast Address Translation

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet- Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Abstract

Network Address Translation is a technique for reducing the complexity of address re-assignment. It can also lead to a reduction in the routing state in the core.

This draft is about using the technique for the same reasons for multicast. The approach is complementary to IP-in-IP tunnels for multicast, but avoids the overhead of encapsulation. It has the same state setup requirements at the edges, as does NAT.

For multicast sessions, there is often an aggregate known as the "cross section" of traffic which is of sufficient common interest that it is carried across the core. In this case, MAT can be used to reduce the multicast olist routing state in the access routers, and the full (S,G) state in the core.

Some promising candidate algorithms for this are described in the paper [NGC2001]. As well as extracting the tree from the routers and

computing the overlaps and so on, clearly, a protocol to establish the state for mapping would be required in exactly the same way that a tunnel setup protocol is needed. This is for further work.

This technique is not necessary for Source Specific Multicast (although it could be used). The interaction with IGMPv3 is to be studied.

1. Introduction

In this document we propose the possible use of address translation as a means to aggregate state for multicast groups that have overlapping coverage in the core.

This type of thing has been proposed before but typically by use of tunnels [Infocom 1998], or implicitly by examination of the possibility of state aggregation merely within a router [Infocom 2000]. Both of these approaches have drawbacks.

The first has the problem with the additional header overhead and subsequent problems for MTU discovery (or configuration). The second is difficult since many service providers wish to account for traffic on a per (S,G, iif/oif) basis, and the aggregation technique potentially masks this. We do not intend to solve this latter problem (if providers want fine grain accounting they have to provision routers for fine grain state). We are proposing something that might avoid the former problem.

Recently, aggregation of state amongst multiple trees was examined for realistic group numbers, membership distribution and network topologies [NGC 2001]. In that work, no specific mechanism other than encapsulation was proposed. As above, this has drawbacks of the additional header overhead.

However, there is no particular reason why one could not use address translation to perform the aggregation, instead of tunnels.

If two trees are congruent in a region, then we can translate all the destination addresses, Gi, into one G (e.g. G0). Usually, sources are distinct for distinct groups. In the case where a source is sending to more than group for which address translation is being performed in a region, we propose translating the source address(es) for the later created (or detected) group(s) at the ingress, and then re-translating at the egress point from the core. (Si to Si' mapping).

A translation management protocol would piggy back on the routing protocol, and would contain for the various base Group (G0), the list of Gi, and the Si to Si' mappings.

To preserve the RPF safety feature of multicast trees, we suggest that the mapping is from Si to an Si', drawn from an equal or more specific prefix that matches.

Where G' has near coverage, but not perfect congruence, add links to egress routers where there are receivers for Gi, but not G0, where additional load is small, and add reverse re-translators back to Si for traffic where there are receivers for G0 but not Gi so that normal filtering removes the traffic (or add a filter - same h/w and s/w support often involved in many router operating systems anyhow).

The rules for "near" coverage could be the same as those used to configure the RP to source tree switch in PIM-SM (or similar traffic threshold derived rules).

An alternative approach to traffic based tree merging, would be to use actual topological similarity. Two approaches suggest themselves

1. Code the tree topologies as a data structure such as a depth then breadth, and do a pure length comparison: if the structures map to a certain threshold point, say they are "equivalent" the threshold to be set by managers at each node in the tree.

2. Use information from neighbouring routers in the tree- e.g. we can split neighbours into "same versus different AS", "same versus different level of OSPF or ISIS", "same versus different prefix", "same prefix up to some point". A Bloom filter could be applied with the "leakiness" deliberately set high (high is a TBD parameter!), which can then allow for approximate matching.

2. MAT and NAT

One view of of address translation [RFC2663] is that it provides a valuable way to conserve addresses. Another view might be that it reduces the globally visible routing state since it means that only active hosts at sites use globally visible prefixes, and thus require forwarding entries. This is, of course, not free - it incurs both the translation state at the edges, and the protocol messages (e.g. via an ALG) to setup and maintain the translation state. In the case of Real Specific IP, the idea is extended to multiple consenting peer realms, rather than between a single global visible address space and any of a set of private address spaces. There is potentially an analogy between Real Specific IP and the use of MATs for Administratively Scoped multicast addresses, which should be explored in the interests of address conservation, as well as the main goal of core router multicast forwarding state reduction.

An important aspect of NATs that has to be understood for Multicast Address Translation is the relationship between packet flows and session flows. In unicast applications to date, most the common cases are client-server, and the TU approach can be used to model how a session initiation relates to the subsequent packet flow, and therefore can be used to trigger (or figure out where to add explicitly) the NAT state instantiation.

Multicast is somewhat different. We need to look at both the receiver oriented nature of multicast (i.e. IGMP, and the prune/graft messages in PIM), and the data driven aspects of some tree building (PIM SM and DVMRP and PIM DM flood and prune).

Finally, higher level protocols that carry IP addresses are also important. In fact, NATs must be aware of them (or else near by firewalls and application layer gateways must coordinate with NATs to translate the application layer messages to, subject to all the security problems of being a "man in the middle", albeit with "permission").

Note that there are already many examples of problems caused by the fact that multicast applications do not rely on the 5-tuple that unicast applications employ, (often because of the liberal use of multiple groups, and also because of other levels of multiplexing). Hence RTP uses CSRC and SSRC fields to indicate the contributer, while some reliable multicast protocols use their own payload multiplex. This means that a simple port-MAP is not simple, or indeed effective (or even feasible in some cases as already discovered for port NATs with unicast RTP and RTCP on two ports!).

In multicast, this means that we must understand SIP and SAP messages, as well as potentially RTSP and several others. The problem here is that we do not currently have the luxury of using the FQDN as a stable namespace to hang the translations from, as NATs do. MATs have to contend with a variety of session end point naming schemes.

Last but not least, in inter-domain multicast, current practice is to use MSDP, which entails source advertisement. However, we envisage a set of MAT servers which coordinate around the edge of a core domain, and can typically be co-located with the MSDP servers in any case, so that the appropriate translations to source advertisements can be easily achieved.

3. Deployment

A key problem with this is deployment. However, we have NAT capable routers in many places. I see this as a minor mod to a family of protocols.

One interesting possibility to consider is that one could multicast the mappings. However, note that this would have to be done reliably (e.g. using PGM or SRM or similar). The same idea has been discussed in terms of scaling BGP (instead of a mesh of TCP connections, use a reliable multicast protocol) and also for link state information flooding.

If we run out of Si's from the same prefix space, fall back to no translation, or else to tunnels. This, obviously should be configurable, as should the thresholds used to set the MAT aggregation.

4. Discussion

This document is a stake in the turf concerning the use of address translation for multicast.

A number of obvious questions need clarifying before the work can be continued (or discontinued).

0. Failure modes (the system should fail safe through appropriate use of soft state and timers).

1. Protocols that mention multicast addresses. We've mentioned some. We are sure there are many others!

2. Feedback - many protocols use feedback, even multicast protocols. For example, reliable multicast transport protocols such as PGM, RLC, and RMTP use feedback messages. Layered coded multicast protocols ("multi-rate congestion control") may have very interesting interactions with the proposed scheme here. Most notably, the aggregation here may make the reverse path trees different again. However, many multicast applications and transport protocols already have to deal with asymmetry (e.g. inter-domain unicast routes through BGP is asymmetric, and non-bi-dir PIM creates asymmetric routes). Note some systems have several levels of feedback (e.g. RTP, RTCP, where ports differ). If a port MAT was required, this would need further consideration.

3. Unicast - One can imagine a case of ultimate aggregation ,where a unicast translation is done. This might enable, for example, multicast islands to communicate via unicast only cores. Some researchers have commented that really fast (partially optical) future IP routers may be hard to build if they have to support packet replication for multicast fan out. Its possible that this scheme could provide a workaround for that.

Note also that one could use a MAT as an alternative to tunnels for edge access from dial-up or other edge networks (e.g. non IGMP/multicast capable DSLAM or Cable Modem access nets), into a multicast capable core. Given these are the same places NATs get deployed, this could easily leverage that.

4. SSM- Source Specific multicast propagates IGMPv3 information. The interaction between that and this needs close examination.

5. The idea of Realm Specific Multicast IP (administrative scoped addresses and MATs) needs further exploration.

etc

5. Acknowledgements

Thanks are due to Colin Perkins for a discussion at NGC in November in London, England.

6. References

[RFC2663] IP Network Address Translator (NAT) Terminology and Considerations, P. Srisuresh, M. Holdrege, Aug 1999.

[RFC3102] Realm Specific IP: Framework, M. Borella, J. Lo, D. Grabelsky, G. Montenegro, October 2001.

[RFC3022] Traditional IP Network Address Translator (Traditional NAT), P. Srisuresh, K. Egevang, Jan 2001.

[Infocom 1998] Forwarding State Reduction for sparse mode multicast communications, J. Tian, G Neufeld, in Proc of IEEE Infocom 1998, March 1998.

[Infocom 2000] On the aggregatability of multicast forwarding state, D. Thaler, M. Handley in Proc of IEEE Infocom 2000, March 2000

[NGC 2001] Aggregated Multicast with Inter-Group Tree Sharing, Aiguo Fei, Junhong Cei, Mario Gerla, Michalis Faloutsos in Proc NGC 2001, November, 2001, available online via http://www.cs.ucla.edu/NRL/hpi/papers.html

7. Security Considerations

The E2E mantra is violated badly.

Black holing attacks are very possible

DDOS on the MATs is clearly quite a possibility.

As usual, all the caveats of intermediate devices that require some information about higher levels apply.

8. IANA Considerations

There are no IANA considerations regarding this document yet. If this note sees any subsequent work, then I would expect at least one protocol to emerge, in which case code points will be needed.

AUTHORS' ADDRESSES

Jon Crowcroft

Marconi Professor of Communications Systems
University of Cambridge
Computer Laboratory
William Gates Building
J J Thomson Avenue
Cambridge
CB3 0FD
UK

E-Mail: Jon.Crowcroft@cl.cam.ac.uk
Tel: +44 (0)1223 763633
Fax: +44 (0)1223 334 678

This draft was created in November 2001.
It expires April 2002.