Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 10, 2015

G. Shepherd
Cisco
A. Dolganow
Alcatel-Lucent
A. Gulko
Thomson Reuters

February 6, 2015

Bit Indexed Explicit Replication (BIER) Problem Statement
draft-shepherd-bier-problem-statement-02

Abstract

There is a need to simplify network operations for multicast services. Current solutions require a tree-building control plane to build and maintain end-to-end tree state per flow, impacting router state capacity and network convergence times. Multi-point tree building protocols are often considered complex to deploy and debug and may include mechanics from legacy use-cases and/or assumptions which no longer apply to the current use-cases. When multicast services are transiting a provider network through an overlay, the core network has a choice to either aggregate customer state into a minimum set of core states resulting in flooding traffic to unwanted network end-points, or to map per-customer, per-flow tree state directly into the provider core state amplifying the network-wide state problem.

Table of Contents

1. Introduction

There is a need to simplify network operations for multicast services. Current solutions require a tree-building control plane, to build and maintain end-to-end tree state per flow, impacting router state capacity and network convergence times. Multi-point tree building protocols are often considered complex to deploy and debug and include mechanics from legacy use-cases and/or assumptions which may no longer apply to the current use-case. When multicast services are transiting a provider network through an overlay, the core network has a choice to either aggregate customer state into a minimum set of core states resulting in flooding traffic to unwanted network end-points, or to map per-customer, per-flow tree state

directly into the provider core state amplifying the network-wide state problem.

This document attempts to discuss the uses, benefits and challenges of the current multicast solutions and to put them in an historical context to better understand why we are where we are today, and to provide a framework for discussion around new solutions that may address our current requirements and challenges.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Objectives

IP Multicast services have been widely adopted in networks where the benefits of efficient, concurrent delivery of content to a sufficiently large set of receivers outweighs the complexity and challenges of deploying and managing the current set of multicast protocols. These deployments are primarily dedicated multicast islands with very little cross-domain inter-networking, and fall short of the early dreams of a multicast enabled Internet.

Multicast began with a large set of requirements shoehorned into a single, complex protocol. Over time, multicast protocols essentially devolved into a set of more simple components to overcome the original complexity, and to address a growing set of use cases. Many of the early complexity can be avoided today by correctly selecting your service model and protocols. But the standard set of protocols available can still be considered overloaded for various reasons.

The current problems associated with the today's multicast solutions can be stated as follows:

- Current multicast methods all require explicit tree building protocols, thereby incurring a lot of state in the transit nodes.

- Receiver driven tree state uses Reverse Path Forwarding (RPF) to build the trees toward the root which often results in multicast forwarding following different paths than unicast forwarding between the same two endpoints.

- Multicast convergence times are negatively impacted by tree state. Any network transition requires unicast to first converge. Once unicast has converged multicast must then recalculate RPF for every tree and rebuild the trees by sending join messages toward

the new RPF neighbor per tree. Joins toward a common RPF neighbor can be aggregated but only up to the link MTU. In large multicast deployments this can result in multicast convergence times of up to a minute or more. In extreme cases the active state may time out before all the new joins are sent and received resulting in multicast to permanently fail after a network failure event even though there is a restored path. This has put an upward bound on the amount of state a multicast network can support.

- Current multicast methods, if they are to provide optimal delivery of multicast packets, require one explicitly built tree per multicast flow; there is no way to aggregate flows (having one state for multiple flows) without sacrificing optimal delivery. In the case of Multicast Virtual Private Network (MVPN) deployments, the operator is forced to choose between unwanted flooded traffic across an aggregate state entry and exposing customer state in the core.

- Some multicast solutions include data-driven events. This has required specialized capabilities to be integrated into routing equipment to protect the control plane from the multicast data plane increasing the cost of multicast support in routing equipment.

- Maintaining and troubleshooting multicast networks can be very difficult. The available solutions are so different than unicast, often revealing unique corner cases that specialized training and skills, and frequently dedicated staff are required just to operate multicast services on a network.

- Current Multicast Virtual Private Networks [RFC6513](MVPN) introduced Border Gateway Protocol[RFC4271](BGP) routes for neighbour discovery and Protocol Independent Multicast[RFC4601] (PIM) Join/Prune propagation. In some deployments when many Multicast MVPNs with many Provider Edge (PE) routers exist in a network and at least some of those MVPNs have a large number of customer-multicast flows, the resulting tax on BGP may be deemed undesired as millions of BGP routes can easily result from multicast deployments. Therefore a solution that allows large MVPN scale with large number of edge PEs and c-multicast flows per MVPN is desired.

- With the introductions of Segment Routing, some networks may elect to remove the Multiprotocol Lable Switching[RFC3031](MPLS) control plane and rely on Interior Gateway Protocol-only or Software Defined Networking-based Segment Routing. In such networks the alternative to existing mechanisms is needed for multicast. Removing the MPLS control plane for unicast makes

little sense unless the multicast control plane also gets simplified.

- The benefits of multi-point services are well understood, but the challenges with the current solutions often result in a failed cost/benefit analysis. Today only those networks with an overwhelming business need have successful multicast deployments, and the rest of the community have come to think of multicast as a failed technology.

How did we get here? What follows is a semi-chronological tour through the devolution of multicast protocols, solutions, and use-cases, describing why earlier complexities and challenges existed, and how they were overcome. This may help frame future work to overcome our current challenges.

3. Deering's Multicast Model

The original Multicast Extensions to the Internet Protocol [RFC0966] and Host Extensions for IP Multicasting [RFC1112]were envisioned by Stephen Deering as part of his graduate work at Stanford University. The need for a multi-point service model was motivated by the advent and deployment of layer3 network topologies breaking existing layer2 applications. The need arose to create an underlay service with the characteristics of a broadcast domain to allow these layer2 applications to continue to function without modification across a layer3 infrastructure.

Though the community quickly saw the value and envisioned many other uses for a multi-point service model, a broadcast domain remained the target model for the solution and the list of requirements focused around those of a broadcast domain. For simplicity the rules of this underlay broadcast domain can be summed up as follows: anyone can send packets into the domain; all members will receive all packets sent into the domain. In order for these layer2 applications to function across this broadcast domain overlay, all of the functions to provide this service were loaded onto network layer.

This new multi-point model was called Multicast. The first multicast solution adopted by the IETF was Distance Vector Multicast Routing Protocol [RFC1075](DVMRP). As the name implies, DVMRP uses a distance-vector routing algorithm derived from Routing Information Protocol [RFC1058](RIP) in combination with the Truncated Reverse Path Broadcasting (TRPB) algorithm to build and maintain tree state and forward multicast packets along these distribution trees. The Internet Assigned Numbers Authority (IANA) was asked to reserve a portion of the global IPv4 address space for multicast destination

addresses required by this model, and in response 224/4 was allocated as the Class D address space for IP Multicast group addresses.

DVMRP has no concept of a "join" message. All new source packets for any given group were simply flooded downstream--essentially broadcasted--following the DVMRP topology. Each leaf of the tree was responsible for sending Non Membership Reports (NMR--prunes) toward the source if there were no downstream receivers for the group. This mechanism came to be known as flood-and-prune, and is a very primitive form of network-based source discovery that all the contemporary applications came to depend on. These contemporary applications were inherently many-to-many either by the nature of the data distribution model, or at the least depended on the many-to-many nature of the network-based-source discovery mechanism.

DVMRP also incorporated the IETF's first specification of an encapsulated overlay. It was clear that this new model would not be supported by every node in the path, and an encapsulation allowed early adopters to build a global multi-point, or multicast capable topology as an overlay.

For clarity of discussion, the functions of the Deering model can be described as:

- Tree building and maintenance

- Network-based source discovery

- Source route information

- Overlay mechanism - tunneling

DVMRP was considered over-loaded in that it carries network source routing information within the protocol in parallel to any existing Interior Gateway Protocol (IGP) generated local routing table. The next generation goal was to focus on the multi-point services needed for the model but to use the local, native routing table as needed for Reverse Path Check (RPF). From this came the advent of Protocol Independent Multicast Sparse Mode [RFC4601](PIM-SM) and Protocol Independent Multicast Dense Mode [RFC3973](PIM-DM). PIM removed any embedded source routing function from the protocol, and instead relied on the exiting routing table as generated from the deployed IGP. PIM also removed any overlay functionality, but retained network-based source discovery as a fundamental part of the protocol. Oops.

4. Network Based Source Discovery

The Deering model introduced the concept of a Group address (G) representing a single broadcast domain. Any source is allowed to send to the group address and the multicast routing infrastructure will build tree state from every source to all interested receivers. All group members only need to signal their G membership to the network and the network will ensure that all source traffic sending to that same group address will arrive at all group members. The network-based source discovery operation providing these functions was intended to provide operational constancy with a layer2 broadcast domain, but comes at significant cost.

Allowing any source to send to a group is an obvious security vulnerability. Many implementations today provide various layers of access control both at the edges and core of the network just to overcome the security concerns for the basic operation of the multicast network.

Network-based source discovery methods can be grouped into two types; flood and prune (DVMRP, PIM-DM), or explicit join (PIM-SM). Both methods depend on the arrival of data to trigger complex network functions to build and maintain the per-source distribution of data for every group. Multicast is often considered complex, fragile, and difficult to troubleshoot, but it is most often the network-based source discovery functions that are the cause of this reputation.

The majority of the use-cases for multicast today are for content with well-know sources. The development of Internet Group Membership Protocol [RFC3376] (IGMPv3) provided a mechanism for group members to signal interest in a source and a group, eliminating the need for network-based source discovery, and facilitating the advent of Source Specific Multicast [RFC4607] (SSM). Many operators still ask how potential SSM group members learn about the sources. The answer is simply to use the same mechanism in which they learned about the group - out-of-band. Source (and group) discovery mechanisms are better served at the application layer for most use-cases. With SSM multicast content can be forwarded and constrained to a single source-rooted tree, or (S,G) channel which has several key benefits:

- Simplified configuration and operation

- Elimination of rouge sources 'stealing' receivers

- Elimination of rouge sources consuming network resources

- Elimination of group address resource restrictions

5. Receiver Driven State

Today's multicast solutions are primarily receiver driven. This is a logical approach in that it is the receiver that decides if and when to join or leave a group or channel. Receiver driven distribution trees built hop-by-hop are an efficient way to dynamically build and scale very large membership fanout. It can be argued that a receiver driven tree's radius can scale infinitely without impact to any upstream segment or node for that tree. But it does then require forwarding state for each tree, or pre-flow state.

The joins propagate upstream from the receiver toward the source or root of the tree, following the unicast routing table. But this reverse path may differ from the optimal unicast forwarding path from the source to the receiver. The result is multicast traffic potentially taking a different forwarding path than unicast traffic between the same to network endpoints. This can often complicate network and traffic engineering.

Each of the existing multicast solutions today, native or overlay, builds and maintains forwarding state per flow, or aggregates some flows into a subset of flow-states. On the surface this may look like an unbounded problem, but in actuality the flow state is only present along the branches of the tree, and no one router needs to maintain global tree state. Router state capacity is not infinite, and this coupling of receiver actions to network state is a potential Denial of Service (DoS) vector. Most implementations today have provided filtering and state-limiting capabilities to secure the multicast infrastructure from this vulnerability.

Increasing multicast forwarding state can also negatively impact network convergence performance. Unicast is only concerned with topology, and any topology changes can converge in a relatively bounded amount of time. The same topology change requires the multicast protocol to rebuild the forwarding state for every active flow. The resulting multicast convergence times are directly dependent on the amount of flow state affected by the convergence event. In extreme cases, the sending, receiving, and processing of the join state for all active flows can exceed the flow state timers resulting in a race condition in which convergence never occurs. Today's implementations have had to incorporate various proprietary solutions to improve network convergence times in large flow-state multicast deployments.

The pros and cons of receiver driven state are as follows:

Pros:

Infinitely scales distribution radius

Aligns with receiver driven join model

Cons:

Potential state DoS vector

Host driven network events

Unbounded per-flow state

Unicast/Multicast traffic divergence

Non-deterministic join latency

Convergence times increasing with flow-state

6. Multicast Virtual Private Networks

Multicast Virtual Private Networks [RFC6513](MVPN) are solutions which allow a core network to transit edge network multicast flows over a core transit network to and from only those MVPN member nodes, without exposing the edge network addressing into the core network forwarding state. The solutions attempt to minimize core state by aggregating trees per-VRF/PE. But this aggregation has the side affect of sending all multicast traffic from that VRF/PE to all other VRF/PE members, whether or not they have down stream flow state.

Various optimizations are available to selectively de-aggregate flow state to better constrain the traffic distribution to only those VRF/PEs with active state. This becomes a trade off between unwanted traffic and an increase in core flow state. These solutions are often data driven resulting in core router state being triggered by date and receiver events.

In addition to a potential BGP route explosion due to an MVPN deployment scale as discussed in section 2, another issue with MVPN relates to architectures used when MVPN deployments require both video-distribution-like model, well served by point-to-multipoint (P2MP) connectivity, and many-to-many model requiring Multipoint-to-multipoint (MP2MP) connectivity. Today, if both models are deployed in a single network, either MP2MP or a mesh of P2MP trees needs to be established, or dual P2MP/MP2MP mLDP architecture may be used, or MP2MP mLDP can be used for both P2MP and MP2MP connectivity. None of those models is optimal as each requires a trade-off between supported protocols, optimal delivery, and operational complexity.

7. Overlay

Deering had the correct insight to assume not every node in a network would be capable of natively transiting multicast flows. The migration to PIM was an attempt to move to a completely native model, which was the right direction. But in this move it also abandoned any other solution for incorporating an underlay into the topology for those portions of the network which for whatever reason do not support native multicast. Early deployments of PIM often incorporated static Generic Routing Encapsulation [RFC2784](GRE) tunnels between PIM domains in an attempt to create an inter domain multicast deployment.

Static tunneling has it's use cases and benefits, but it is not the ideal tool to dynamically stitch together a large and topologically diverse receiver population. A receiver driven distribution model would be better served with a receiver driving overlay mechanism. This would indicate that when overlay was removed from the tree building protocol it should have migrated to IGMPv3 and Multicast Listener Discovery [RFC3810](MLDv2), the membership protocol, but it was seen as a necessary requirement at that time. To fill this requirement today Automatic Multicast Tunnels (AMT) is being progressed as the overlay standard for bridging multicast interested receivers over unicast only intermediate networks.

8. Summary

Multicast began with a heavily overloaded protocol DVMRP, and has evolved over time by removing functionality from this all-in-one solution, and off-loading certain function to either more specialized protocols or existing protocols and functions. Multicast has what may be the unique distinction of starting very complex, but evolving through more simple stages along the way. It may be time to consider the next step in the evolution toward simplicity.

Today we depend on receiver driven joins propagating end-to-end from receivers toward sources, and maintaining per-flow state in every node along the path. This state crosses administrative domains. Unicast has a simple model where local specificity stays local and does not directly impact the global table. Multicast state has no administrative boundaries today. It may be beneficial to consider the autonomy of networks in the path, and their specific topology and requirements. PIM successfully utilizes the available routing table for RPF checks and joins. This route table may also be considered as a source of topology information for a set of receiver nodes within a given network.

9. IANA Considerations

This memo includes no request to IANA.

All drafts are required to have an IANA considerations section (see Guidelines for Writing an IANA Considerations Section in RFCs [RFC5226] for a guide). If the draft does not require IANA to do anything, the section contains an explicit statement that this is the case (as above). If there are no requirements for IANA, the section will be removed during conversion into an RFC by the RFC Editor.

10. Security Considerations

All drafts are required to have a security considerations section. See RFC 3552 [RFC3552] for a guide.

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

11.2. Informative References

[RFC0966] Deering, S. and D. Cheriton, "Host groups: A multicast extension to the Internet Protocol", RFC 966, December 1985.

[RFC1058] Hedrick, C., "Routing Information Protocol", RFC 1058, June 1988.

[RFC1075] Waitzman, D., Partridge, C., and S. Deering, "Distance Vector Multicast Routing Protocol", RFC 1075, November 1988.

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, August 1989.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.

[RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.

[RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, July 2003.

[RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.

[RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, January 2005.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

Appendix A. Additional Stuff

This becomes an Appendix.

Authors' Addresses

Greg Shepherd (editor)
Cisco
170 W. Tasman Dr.
San Jose
US

Email: gjshep@gmail.com

Andrew Dolganow (editor)
Alcatel-Lucent
600 March Rd.
Ottawa, Ontario K2K 2E6
Canada

Email: andrew.dolganow@alcatel-lucent.com

Arkadiy Gulko (editor)
Thomson Reuters

Email: arkadiy.gulko@thomsonreuters.com