RTGWG Working Group
Internet-Draft
Intended Status: Experimental RFC

Expires: September 2012

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
Vasan Srini
I.I.T Madras

March 27, 2012

Building power optimal Multicast Trees
draft-mjsraman-rtgwg-pim-power-02

Abstract

Power consumption in multicast replication operations is an area of concern and choosing suitable replication points that can decrease power consumption overall assumes importance. Multicast replication capacity is an attribute of every line card of major routers and multi-layer switches that support multicast in the core of an Internet Service Provider (ISP) or an enterprise network.

Currently multicast replication points on Point-to-Multipoint Multicast Distribution trees consume power while delivering multiple output streams of data from a given input stream. The multicast distribution trees are constructed without any regard for a proper placement of the replication points and consequent optimal power consumption at these points.

This results in overloading certain routers while under-utilizing others. An optimal usage of these replication resources could reduce power consumption on these routers bringing power consumption to optimality. In this paper, we propose a mechanism by which Multicast Distribution Trees are constructed for carrying multicast traffic across multiple routers within a given network. We propose that these Multicast Distribution Trees be built by using the information pertaining to power-replication capacity ratio available with fine grained components such as multicast capable line-cards of routers and multi-layer switches deployed within a network.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

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

Copyright and License Notice

Table of Contents

1 Introduction

Multicast traffic across multiple areas within a given network such as an ISP or a Campus Environment Network, may be carried using Multicast Distribution Trees. The traffic may be carried from a ingress router to several egress routers, example in a Campus Environment network. The Network under consideration may comprise of multiple areas involving a backbone area and several non-backbone areas connected to each other through the backbone. If several such multicast streams are to be carried in the network, it would be most useful to have such Multicast Distribution Trees constructed such that they have optimal power to available replication capacity ratios on the routers' linecards that they traverse from source to destinations. The intent is to provide a solution whereby several such Distribution Trees can be laid out in such a way that the set of routers that replicate multicast traffic traversed by the trees are most optimal in the utilization of the power provided to them given that there is sufficient replication capacity available. This we believe would essentially lead to a equilibrium of power to available replication capacity ratios amongst all routers in the topology which in turn would optimize and reduce the overall ratios for the network.

Each router and its respective linecards deployed in the network have an advertised capability for replication. Most multi-layer switches and routers from vendors advertise in their respective data sheets a certain capability for replication for each type of linecard deployable on the box. Replication consumes power and delivers multiple streams of data from a given input stream. It is status quo that (Point-to-Multipoint) P2MP trees are constructed without taking into account the power to available replication capacity ratios of such routers thus overloading certain routers while underutilizing the others. An optimal usage of these resources could reduce power consumption on these routers / multi-layer switches. This equilibrium could be arrived at by using a capability to choose from each downstream PIM router the most power optimal path to the selected (through current mechanisms) PIM upstream neighbor in the PIM-based Multicast Distribution Tree which may be a shared tree or a Shortest Path Tree as the case may be. The metric used to select the upstream PIM neighbor could be the power to available replication capacity ratio of each of the said router's line cards that are part of the ECMP set of paths to the upstream neighbor if such ECMP paths do exist. The metric comparison is done for all ECMP paths and the line cards involved therein depending on their current utilization of their replication capacity and power consumption.

This paper is organized as follows; In section 2, we deal with the scheme that we propose. In section 2.1, we discuss some examples of the scheme at work, and in section 3 we conclude with further areas

of study that may be useful to undertake.

1.1 Terminology

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. Methodology of the proposal

The key metric under consideration is the power consumed DIVIDED BY available replication capacity on each of the linecards of a router in the network whose constituent ports form part of a ECMP set of paths to a PIM upstream neighbor. The said ports on the different line cards that form the ECMP set of links are eligible to be used as a linecard:port atop which multicast traffic on that tree can be carried. When choosing the path from a ECMP set of paths to a PIM upstream neighbor, the said downstream PIM neighbor calculates the power to multicast replication capacity ratio for each of the line cards that are eligible to be chosen as the linecard:port combination to be used in that section of the distribution tree. The lowest ratio decides which linecard is chosen and if there exist multiple ports within that linecard that connect to the said PIM upstream neighbor the usual algorithm is used to select one of those ports. The key proposal that this document recommends is the use of the power-multicast-replication-capacity ratio to choose from among the different linecards. The choice of port is left to the standard method.

Assume that the following router topology in the vicinity of the sender / senders is computed.

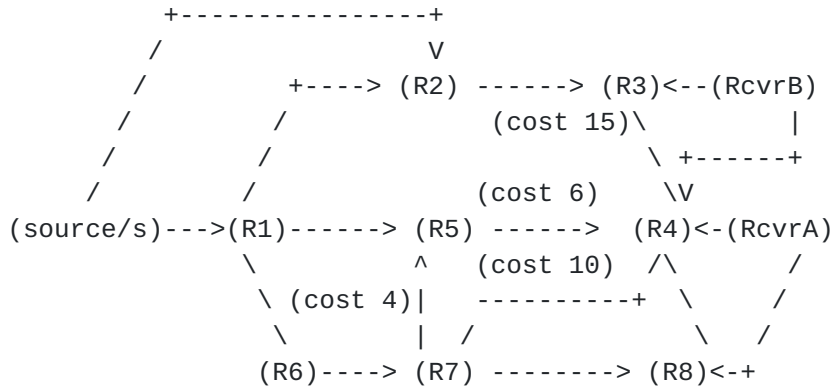
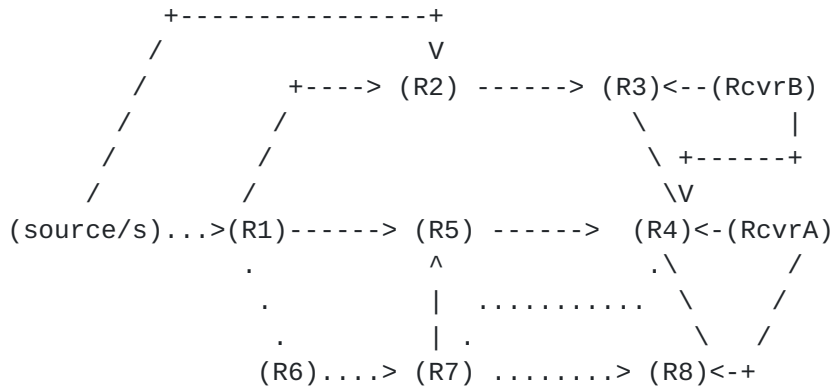


Figure 1: Topology within a given network with an upstream ECMP link from R4 to R7

In the above diagram you can see that the source/sources are connected using a multi homed connections to the same ISP through Routers R1 nd R2. Similarly there are two Receiver sites RcvrA and RcvrB that are multihomed to TWO Routers RcvrB to R3 and R4 and for RcvrA to R4 and R8 respectively. You can also observe that R4 is connected to R7 through multiple paths. Assuming that both these paths are Equal Cost then this gives rise to a situation where ECMP paths exist for the PIM downstream router R4 to the PIM upstream router R7.

Consider that RcvrA sends an IGMP join to R4. R4 now needs to send a PIM join towards the upstream router R7. Assume this is a shared tree with Rendezvous Point (RP) as R7. There are 2 equal cost paths to R7 from R4 each with cost 10 ((R7->R5->R4 = 6 + 4 = 10) and (R7 -> R4 = 10)). Assume that each of these paths from R4 to R5 onto R7 and from R4 to R7 directly are on different linecards in the chassis R4. Normally one of them would be chosen and power-to-multicast-replication-capacity would not be a consideration in that decision. What this document proposes is that R4 consider the metric PWR which is a ratio formed by dividing the power consumed on each of the linecards by their respective current multicast replication capacity.

Obviously one of them would have to be chosen. In the metric comparison the linecard that has the lower PWR metric wins and is selected for consideration to send a PIM join to R7 (the PIM upstream neighbor and in this case the RP as well).



Legend : dotted lines represent path computed.

Figure 2: Instantiating an optimal power consuming distribution tree

In our example as in Figure 2 we find that the direct link to R4 and R7 wins out as the link to be used in the distribution tree.

The one exception that SHOULD be considered in this decision is that if the Outgoing Interface List consists of ports on linecard X on

which R4's downstream PIM neighbors have sent their respective PIM joins and if the ECMP set of paths to the router R7 consist of linecard X and Y, it would be preferable to choose linecard X without taking into consideration the PWR metric. This is in light of the fact that if majority of the OIF list's port members lie on linecard X and the ingress port were also to be placed on linecard X then the replication would be more optimal as it would not have to traverse say the switch fabric to get to the majority of the OIF list. Other localization conditions could also be considered as exceptions to the PWR metric based rule.

This document assumes that the power used by each linecard and the multicast replication utilization and advertised capacity are available as data readable from the hardware on the router chassis under consideration. Please note that unicast traffic already being carried on the linecard may also contribute to the power being consumed at the router's linecards under consideration.

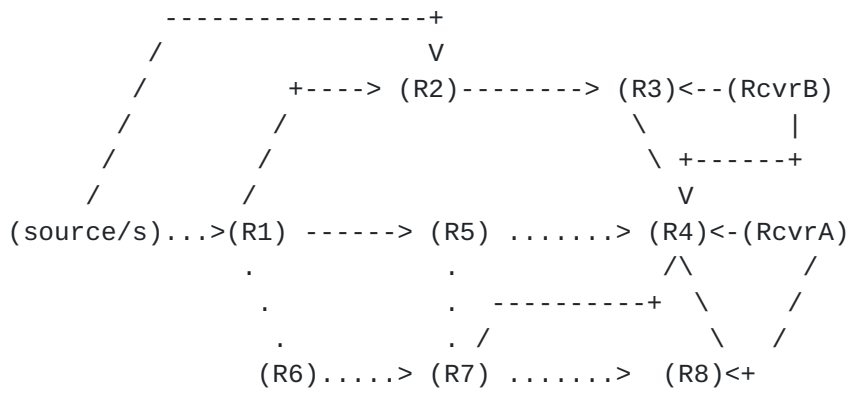
If ECMP paths dont exist then there is no choice to make hence the default selection of the link to be used to send a PIM join to the upstream neighbor is followed.

As a result of this decision to include the PWR metric the paths in the tree where ECMP links occur have the least power to available replication capacity ratios at the time of computation.

Assume the following path is computed as per the least power to available replication capacity ratios. Paths are computed through R6, R7, R8, R4, and say the multicast stream occupies 4GB of traffic along this tree so constructed and the available capacity of these routers reduces to 6GB assuming all of them have a base capacity of 10GB. Subsequent paths constructed would have to take into account the newly computed power to current replication capacity ratio in the topology for multicast streams / trees yet to come. Now the linecard connecting R4 to R7 directly will have reduction of a quantum of 4GB capacity. It would reduce to 6GB as its available capacity.

Assume another 6GB worth of traffic is loaded onto this topology in terms of a multicast stream / multiple streams then the new path computed for these new streams would NOT possibly utilize the same path as computed before since the power utilization and the available replication capacity would have been changed to create a higher PWR ratio. If the old streams reduce the replication capacity to an extent such that routers through which they pass can no longer be used since these routers' power to available replication capacity has become poor when compared to other paths then a different path may be computed from the ingress router to the egress router in such a way as to avoid those routers which have such poor ratios. This again

applies only in ECMP sections of the distribution tree.



Legend : dotted lines represent path computed.

Figure 3: Instantiating a subsequent optimal power consuming distribution tree

Here R4 would now have to choose the path to R7 (which is also the RP) through R5 since the PWR metric on R4 to R7 direct link would have increased as a result of carrying the old stream.

Dynamism in multicast trees is another important point to consider as PIM-Prunes and other PIM-joins may happen with respect to the replication point under consideration. Suitable modifications to the algorithm may be proposed to take into consideration such dynamic conditions without causing major interruption to the multicast flows.

2.1 Discussion of this scheme

This scheme applies to PIM-SM, PIM-SSM. Applicability to PIM-Bidir is also possible but currently not discussed in this document in detail.

Routers may have step levels in which they increase power consumption when they additively are loaded with more large bandwidth consuming multicast streams. Calibrating these levels may be useful for implementing this scheme. It is possible that such calibrated thresholds can be used for calculating the power to available replication capacity ratios in the Multicast environments. This would be useful for bringing down the frequency of calculations on a line-card about its ratios. When power consumption meanders within a certain given interval these ratios need not be calculated even if further multicast streams are added to it. The incentive is to recognize a linecard that does not drastically change power consumption even if large bandwidth streams are added onto it for replication and thus give it credit for its power optimal

functioning. If a linecard on a router tends to consume the highest level of power even when carrying low amounts of multicast streams and replicating them on its line card, it would automatically have a poor ratio when compared to a linecard that efficiently uses power when considering the replication capacity being used. The best case would be a low power consuming line-card or a router filled with such line cards that does not leave its power interval no matter how much ever replication capacity is sought to be used on it. But that would be an ideal condition but it is definitely an idealistic scenario towards which the router manufacturers should look at.

2.2 Pseudo code for the proposed changes

```
If (there exist ECMP paths to a PIM upstream NBR)
     AND (No localized conditions exist)
then
             Calculate PWR ratio for each LC;
             PWR per LC = power consumed by LC /
                             AvailableMCastReplicCap;
             Choose the Lowest PWR;
             Select that LC for the link to send PIM Join;
Endif
```

2.3 Port Choice on same Linecard

In case in the set of ECMP links to the upstream PIM NBR there exist ports from the same line card and there is a tie breaking mechanism required amongst these ports the following changes are recommended.

```
If (there exist ports on the same linecard which
    constitute ECMP paths to a PIM upstream NBR)
     AND (No localized conditions exist)
then
     Choose the Lowest Utilized port;
     Select that port in LC for the link to send PIM Join;
Endif
```

3 Conclusion

Here we propose a scheme that takes into account the power to available replication capacity ratios as weights for the edges which are the ECMP set of paths to a PIM upstream neighbor and compute a low cost power path for multicast replication. This is an area of future study which would be most conducive in terms of bringing about optimal power usage and thus incentivising vendors to manufacture low power consuming equipment. Compelled to bring about radical change in the thinking relating to power consumption vendors manufacturing

networking equipment will drive down power consumption since the scheme proposed chooses or gives priority to low power guzzling linecards.

3 Security Considerations

None.

4 IANA Considerations

None.

5 References

5.1 Normative References

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

[RFC1776] Crocker, S., "The Address is the Message", RFC 1776, April 1 1995.

[TRUTHS] Callon, R., "The Twelve Networking Truths", RFC 1925, April 1 1996.

5.2 Informative References

[1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.

[2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.

[3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1993.

[4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.

[5] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1995.

[6] W. Lu and S. Sahni, Low-power TCAMs for very large

forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.

[7] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.

[EVILBIT] Bellovin, S., "The Security Flag in the IPv4 Header", RFC 3514, April 1 2003.

[RFC5513] Farrel, A., "IANA Considerations for Three Letter Acronyms", RFC 5513, April 1 2009.

[RFC5514] Vyncke, E., "IPv6 over Social Networks", RFC 5514, April 1 2009.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
I.I.T Madras,
Chennai - 600036
TamilNadu,
India.

EMail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,

India.

EMail: gaurav@ee.iitm.ac.in

Vasan Srini,
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

Email: vasan.vs@gmail.com