

TRILL Working Group
Internet Draft
Intended status: Standard Track
Updates: 6325

Tissa Senevirathne
CISCO
Janardhanan Pathangi
DELL
Jon Hudson
Brocade

April 1, 2014

Expires: October 2014

Coordinated Multicast Trees (CMT) for TRILL
draft-ietf-trill-cmt-03.txt

Status of this Memo

This Internet-Draft is submitted 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 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/lid-abstracts.txt>

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

This Internet-Draft will expire on September 1, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

TRILL facilitates loop free connectivity to non-TRILL legacy networks via choice of an Appointed Forwarder for a set of VLANs. Appointed Forwarder provides load sharing based on VLAN with an active-standby model. Mission critical operations such as High Performance Data Centers require active-active load sharing model. The Active-Active load sharing model can be accomplished by representing any given non-TRILL legacy network with a single virtual RBridge. Virtual representation of the non-TRILL legacy network with a single RBridge poses serious challenges in multi-destination RPF (Reverse Path Forwarding) check calculations. This document specifies required enhancements to build Coordinated Multicast Trees (CMT) within the TRILL campus to solve related RPF issues. CMT provides flexibility to RBridges in selecting desired path of association to a given TRILL multi-destination distribution tree.

Table of Contents

1. Introduction.....	3
1.1. Scope and Applicability.....	5
1.2. Contributors.....	5
2. Conventions used in this document.....	5
2.1. Acronyms.....	5
3. The AFFINITY sub-TLV.....	6
4. Multicast Tree Construction and Use of Affinity Sub-TLV.....	6
4.1. Update to RFC 6325.....	7
4.2. Announcing virtual RBridge nickname.....	8
4.3. Affinity Sub-TLV Capability.....	8
5. Theory of operation.....	9
5.1. Distribution Tree provisioning.....	9
5.2. Affinity Sub-TLV advertisement.....	9
5.3. Affinity sub-TLV conflict resolution.....	9
5.4. Ingress Multi-Destination Forwarding.....	10
5.4.1. Forwarding when $n < k$	10
5.5. Egress Multi-Destination Forwarding.....	11
5.5.1. Traffic Arriving on an assigned Tree to RBk-RBv.....	11
5.5.2. Traffic Arriving on other Trees.....	11
5.6. Failure scenarios.....	11

5.6.1. Edge RBridge RBk failure.....	11
5.7. Backward compatibility.....	12
6. Security Considerations.....	12
7. IANA Considerations.....	13
8. References.....	13
8.1. Normative References.....	13
8.2. Informative References.....	14
9. Acknowledgments.....	14
Appendix A. Change History.....	15

1. Introduction

TRILL (Transparent Interconnection of Lots of Links) presented in [RFC6325] and other related documents, provides methods of utilizing all available paths for active forwarding, with minimum configuration. TRILL utilizes IS-IS (Intermediate System to Intermediate System [IS-IS]) as its control plane and uses a TRILL header with hop count.

[RFC6325], [6327bis] and [RFC6439] provide methods for interoperability between TRILL and Legacy networks. [RFC6439], provide an active-standby solution, where only one of the RBridges on a link with end stations is in the active forwarding state for end station traffic for any given VLAN. That RBridge is referred to as the Appointed Forwarder (AF). All frames ingressed into a TRILL network via the Appointed Forwarder are encapsulated with the TRILL header with a nickname held by the ingress AF RBridge. Due to failures, re-configurations and other network dynamics, the Appointed Forwarder for any set of VLANs may change. RBridges maintain forwarding tables that contain destination MAC address and VLAN to egress RBridge binding. In the event of AF change, forwarding tables of remote RBridges may continue to forward traffic to the previous AF and that traffic may get discarded at the egress, causing traffic disruption.

Mission critical applications such as High Performance Data Centers require resiliency during failover. The active-active forwarding model minimizes impact during failures and maximizes the available network bandwidth. A typical deployment scenario, depicted in Figure 1, which may have either End Stations and/or Legacy bridges attached to the RBridges. These Legacy devices typically are multi-homed to several RBridges and treat all of the uplinks as a single Multi-Chassis Link Aggregation (MC-LAG) bundle. The Appointed Forwarder designation presented in [RFC6439] requires each of the edge RBridges to exchange TRILL hello packets. By design, an MC-LAG does not forward packets received on one of the member ports of the MC-LAG to other member ports of the same MC-LAG. As a result the AF

designation methods presented in [RFC6439] cannot be applied to deployment scenario depicted in Figure 1.

An active-active load-sharing model can be implemented by representing the edge of the network connected to a specific edge group of R Bridges by a single virtual R Bridge. Each virtual R Bridge MUST have a nickname unique within its TRILL campus. In addition to an active-active forwarding model, there may be other applications that may require similar representations.

Sections 4.5.1 and 4.5.2 of [RFC6325] as updated by [clearcor] specify distribution tree calculation and RPF (Reverse Path Forwarding) check calculation algorithms for multi-destination forwarding. These algorithms strictly depend on link cost and parent R Bridge priority. As a result, based on the network topology, it may be possible that a given edge R Bridge, if it is forwarding on behalf of the virtual R Bridge, may not have a candidate multicast tree that the edge R Bridge can forward traffic on because there is no tree for which the virtual R Bridge is a leaf node from the edge R Bridge.

In this document we present a method that allows R Bridges to specify the path of association for real or virtual child nodes to distribution trees. Remote R Bridges calculate their forwarding tables and derive the RPF for distribution trees based on the distribution tree association advertisements. In the absence of distribution tree association advertisements, remote R Bridges derive the SPF (Shortest Path First) based on the algorithm specified in section 4.5.1 of [RFC 6325].

Other applications, beside the above mentioned active-active forwarding model, may utilize the distribution tree association framework presented in this document to associate to distribution trees through a preferred path.

This proposal requires presence of multiple multi-destination trees within the TRILL campus and updating all the R Bridges in the network to support the new Affinity sub-TLV (Section 3.). It is expected that both of these requirements will be met as they are control plane changes, and will be common deployment scenarios. In case either of the above two conditions are not met R Bridges MUST support a fallback option for interoperability. Since the fallback is expected to be a temporary phenomenon till all R Bridges are upgraded, this proposal gives guidelines for such fallbacks, and does not mandate or specify any specific set of fallback options.

1.1. Scope and Applicability

This document specifies an Affinity sub-TLV to solve associated RPF issues at the active-active edge. Specific methods in this document for making use of the Affinity sub-TLV are applicable where multiple RBridges are connected to an edge device through multi-chassis link aggregation or to a multiport server or some similar arrangement where the RBridges cannot see each other's Hellos.

This document DOES NOT provide other required operational elements to implement active-active edge solution, such as methods of multi-chassis link aggregation. Solution specific operational elements are outside the scope of this document and will be covered in solution specific documents. (See, for example [TRILLPN].)

Examples provided in this document are for illustration purposes only.

1.2. Contributors

The work in this document is a result of much passionate discussions and contributions from following individuals. Their names are listed in alphabetical order:

Ayan Banerjee, Dinesh Dutt, Donald Eastlake, Mingui Zhang, Radia Perlman, Sam Aldrin, Shivakumar Sundaram and Zhai Hongjun.

2. Conventions used in this document

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 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying [RFC2119] significance.

2.1. Acronyms

MC-LAG: . Multi-Chassis Link Aggregation is a solution specific extension to [8021AX], that facilitates connecting group of links from an originating device (A) to a group of discrete devices (B). Device (A) treats, all of the links in a given Multi-Chassis Link Aggregation bundle as a single logical interface and treats all devices in Group (B) as a single logical device for all forwarding purposes. Device (A) does not forward packets receive on Multi-

Chassis Link bundle out of the same Multi-Chassis link bundle. Figure 1 depicts a specific use case example.

CE : Classical Ethernet device, that is a device that performs forwarding based on 802.1Q bridging. This also can be end-station or a server.

RPF: Reverse Path Forwarding. See section 4.5.2 of [RFC6325].

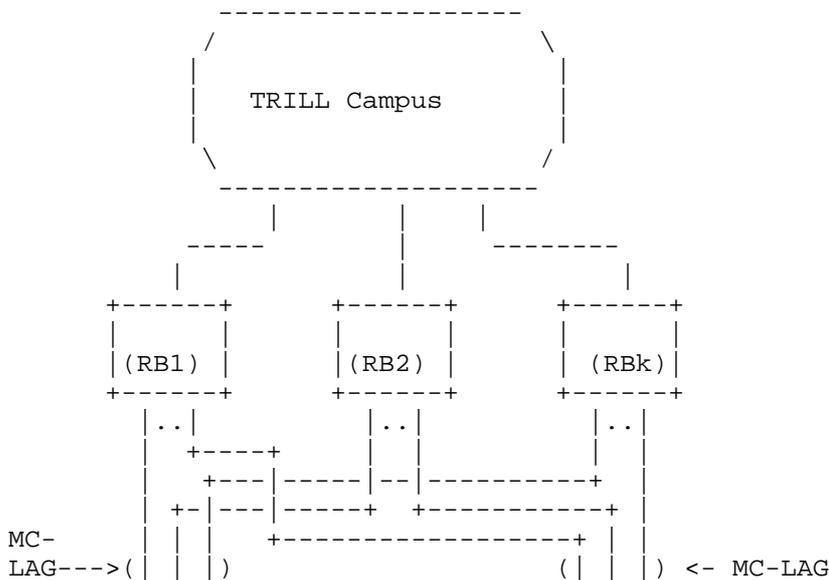
3. The AFFINITY sub-TLV

Association of an RBridge to a multi-destination distribution tree through a specific path is accomplished by using a new IS-IS sub-TLV, the Affinity sub-TLV.

The AFFINITY sub-TLV appears in Router capability TLVs that are within LSP PDUs, as described in [6326bis] which specifies the code point and data structure for the Affinity sub-TLV.

4. Multicast Tree Construction and Use of Affinity Sub-TLV

Figure 1 and Figure 2 below show the reference topology and a logical topology using CMT to provide active-active service.



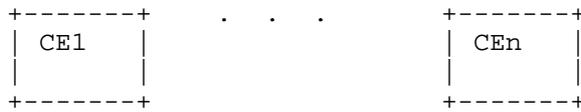


Figure 1 Reference Topology

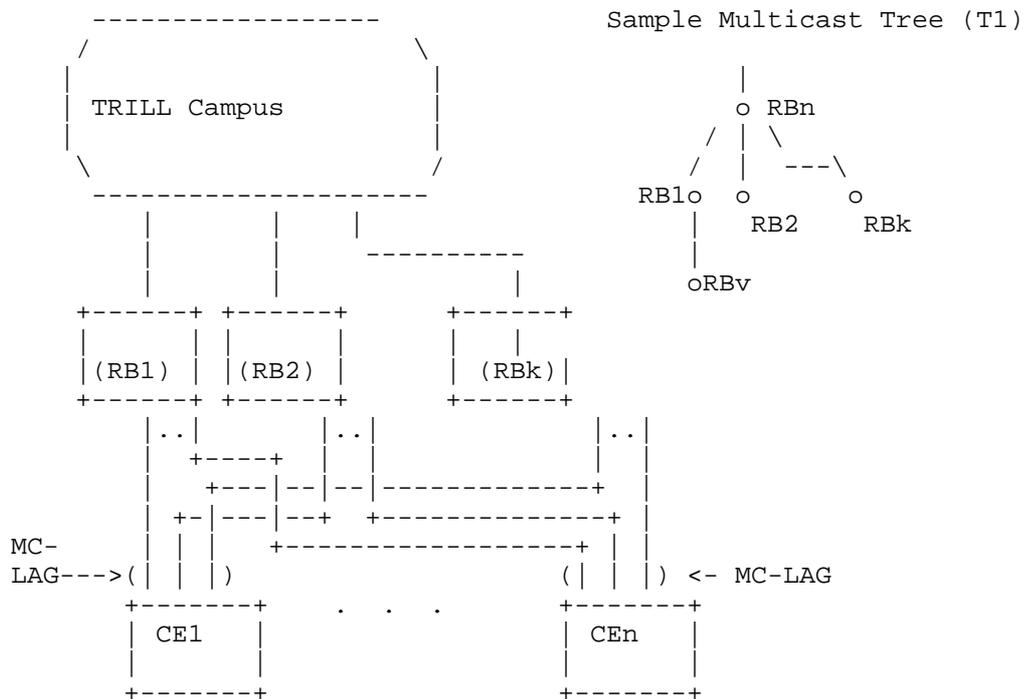


Figure 2 Example Logical Topology

4.1. Update to RFC 6325

Section 4.5.1 of [RFC6325], is updated as below:

Each RBridge that desires to be the parent RBridge for child Rbridge RBy in a multi-destination distribution tree x announces the desired association using an Affinity sub-TLV. The child RBridge RBy is specified by its nickname (or one of its nicknames if it holds more than one).

When such an Affinity sub-TLV is present, the association specified by the affinity sub-TLV MUST be used when constructing the multi destination distribution tree except in case of conflicting Affinity sub-TLV which are resolved as specified in Section 5.3. In the absence of such an Affinity sub-TLV, or if there are any RBridges in the campus that are do not support Affinity sub-TLV, distribution trees tree are calculated as specified in the section 4.5.1 of [RFC6325] as updated by [clearcor]. Section 4.3. below specifies how to identify RBridges that support Affinity sub-TLV capability.

4.2. Announcing virtual RBridge nickname

Each edge RBridge RB1 to RBk advertises in its LSP virtual RBridge nickname RBv using the Nickname sub-TLV (6), [6326bis], along with their regular nickname or nicknames.

It will be possible for any RBridge to determine that RBv is a virtual RBridge because each RBridge (RB1 to RBk) this appears to be advertising that it is holding RBv is also advertising an Affinity sub-TLV asking that RBv be its child in one or more trees.

Virtual RBridges are ignored when determining the distribution tree roots for the campus.

All RBridges outside the edge group assume that multi-destination packets with ingress nickname RBv might use any of the distribution trees that any member of the edge group is advertising that it might use.

4.3. Affinity Sub-TLV Capability.

RBridges that announce the TRILL version sub-TLV [6326bis] and set the Affinity capability bit (Section 7.) support the Affinity sub-TLV and calculation of multi-destination distribution trees and RPF checks as specified herein.

5. Theory of operation

5.1. Distribution Tree provisioning

Let's assume there are n distribution trees and k edge RBridges in the edge group of interest.

If $n \geq k$

Let's assume edge RBridges are sorted in numerically ascending order by SystemID such that $RB1 < RB2 < RBk$. Each Rbridge in the numerically sorted list is assigned a monotonically increasing number j such that; $RB1=0$, $RB2=1$, $RBi=j$ and $RBi+1=j+1$.

Assign each tree to RBi such that tree number $\{ (tree_number) \% k\}+1$ is assigned to RBridge i for tree_number from 1 to n . where n is the number of trees and k is the number of RBridges considered for tree allocation.

If $n < k$

Distribution trees are assigned to RBridges $RB1$ to RBn , using the same algorithm as $n \geq k$ case. RBridges $RBn+1$ to RBk do not participate in active-active forwarding process on behalf of RBv .

5.2. Affinity Sub-TLV advertisement

Each RBridge in the $RB1..RBk$ domain advertises an Affinity TLV for RBv to be its child.

As an example, let's assume that $RB1$ has chosen Trees $t1$ and $tk+1$ on behalf of RBv .

$RB1$ advertises affinity TLV; $\{RBv, Num\ of\ Trees=2, t1, tk+1\}$.

Other RBridges in the $RB1..RBk$ edge group follow the same procedure.

5.3. Affinity sub-TLV conflict resolution

In TRILL, multi-destination distribution trees are built outward from the root. If an RBridge $RB1$ advertises an Affinity sub-TLV with an AFFINITY RECORD that asks for RBridge $RBroot$ to be its child in a tree rooted at $RBroot$, that AFFINITY RECORD is in conflict with TRILL distribution tree root determination and MUST be ignored.

If an RBridge RB1 advertises an Affinity sub-TLV with an AFFINITY RECORD that's ask for nickname RBn to be its child in any tree and RB1 is not adjacent to a real or virtual RBridge RBn, that AFFINITY RECORD is in conflict with the campus topology and MUST be ignored.

If different RBridges advertise Affinity sub-TLVs that try to associate the same virtual RBridge as their child in the same tree or trees, those Affinity sub-TLVs are in conflict for those trees. The nicknames of the conflicting RBridges are compared to identify which RBridge holds the nickname that is the highest priority to be a tree root, with the System ID as the tie breaker

The RBridge with the highest priority to be a tree root will retain the Affinity association. Other RBridges with lower priority to be a tree root MUST stop advertising their conflicting Affinity sub-TLV, re-calculate the multicast tree affinity allocation, and, if appropriate, advertise a new non-conflict Affinity sub-TLV.

Similarly, remote RBridges MUST honor the Affinity sub-TLV from the RBridge with the highest priority to be a tree root (use system-ID as the tie-breaker in the event of conflicting priorities) and ignore the conflicting Affinity sub-TLV entries advertised by the RBridges with lower priorities to be tree roots.

5.4. Ingress Multi-Destination Forwarding

If there is at least one tree on which RBv has affinity via RBk, then RBk performs the following operations, for multi-destination frames received from a CE node:

1. Flood to locally attached CE nodes subjected to VLAN and multicast pruning.
2. Ingress in the TRILL header and assign ingress RBridge nickname as RBv. (nickname of the virtual RBridge).
3. Forward to one of the distribution trees, tree x in which RBv is associated with RBk

5.4.1. Forwarding when $n < k$

If there is no tree on which RBv can claim affinity via RBk (Probably because the number of trees n built is less than number of RBridges k announcing the affinity sub-TLV), then RBk MUST fall back to one of the following

1. This RBridge should stop forwarding frames from the CE nodes, and should mark that port as disabled. This will prevent CE

nodes from forwarding data on to this RBridge, and only use those RBridges which have been assigned a tree -
-OR-

2. This RBridge tunnels multi-destination frames received from attached native devices to an RBridge RBy that has an assigned tree. The tunnel destination should forward it to the TRILL network, and also to its local access links. (The mechanism of tunneling and handshake between the tunnel source and destination are out of scope of this specification and may be addressed in future documents).

Above fallback options may be specific to active-active forwarding scenario. However, as stated above, Affinity sub-TLV may be used in other applications. In such event the application SHOULD specify applicable fallback options.

5.5. Egress Multi-Destination Forwarding

5.5.1. Traffic Arriving on an assigned Tree to RBk-RBv

Multi-destination frames arriving at RBk on a Tree x, where RBk has announced the affinity of RBv via x, MUST be forwarded to CE members of RBv that are in the frame's VLAN. Forwarding to other end-nodes and RBridges that are not part of the network represented by the RBv virtual RBridge MUST follow the forwarding rules specified in [RFC6325].

5.5.2. Traffic Arriving on other Trees

Multi-destination frames arriving at RBk on a Tree y, where RBk has not announced the affinity of RBv via y, MUST NOT be forwarded to CE members of RBv. Forwarding to other end-nodes and RBridges that are not part of the network represented by the RBv virtual RBridge MUST follow the forwarding rules specified in RFC6325.

5.6. Failure scenarios

The below failure recovery algorithm is presented only as a guideline. Implementations MAY include other failure recover algorithms. Details of such algorithms are outside the scope of this document.

5.6.1. Edge RBridge RBk failure

Each of the member RBridges of given virtual RBridge edge group is aware of its member RBridges through configuration or some other method.

Member RBridges detect nodal failure of a member RBridge through IS-IS LSP advertisements or lack thereof.

Upon detecting a member failure, each of the member RBridges of the RBv edge group start recovery timer T_{rec} for failed RBridge R_{Bi} . If the previously failed RBridge R_{Bi} has not recovered after the expiry of timer T_{rec} , members RBridges perform distribution tree assignment algorithm specified in section 5.1. Each of the member RBridges re-advertises the Affinity sub-TLV with new tree assignment. This action causes the campus to update the tree calculation with the new assignment.

R_{Bi} upon start-up, starts advertising its presence through IS-IS LSPs and starts a timer T_i . Member RBridges detecting the presence of R_{Bi} start a timer T_j . Timer T_j SHOULD be at least $< T_i/2$. (Please see note below)

Upon expiry of timer T_j , member RBridges recalculate the multi-destination tree assignment and advertised the related trees using Affinity sub-TLV.

Upon expiry of timer T_i , R_{Bi} recalculate the multi-destination tree assignment and advertises the related trees using Affinity TLV.

Note: Timers T_i and T_j are designed so as to minimize traffic down time and avoid multi-destination packet duplication.

5.7. Backward compatibility

Implementations MUST support backward compatibility mode to interoperate with pre Affinity sub-TLV RBridges in the network. Such backward compatibility operation MAY include, however is not limited to, tunneling and/or active-standby modes of operations.

Example:

Step 1. Stop using virtual RBridge nickname for traffic ingressing from CE nodes

Step 2. Stop performing active-active forwarding. And fall back to active standby forwarding, based on locally defined policies. Definition of such policies is outside the scope of this document and may be addressed in future documents.

6. Security Considerations

In general, the RBridges in a campus are trusted routers and the authenticity of their link state information (LSPs) and link local

PDU (Hello, etc.) can be enforced using regular IS-IS security mechanisms [IS-IS] [RFC5410]. This including authenticating the contents of the PDUs used to transport Affinity sub-TLVs.

The particular Security Considerations involve with different applications of the Affinity sub-TLV will be covered in the document(s) specifying those applications.

For general TRILL Security Considerations, see [RFC6325].

7. IANA Considerations

IANA is requested to allocate a capability bit for "Affinity Supported" in the TRILL-VER sub-TLV. "Affinity Supported" capability bit and Affinity sub-TLV are specified and allocated in [6326bis].

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5310] Bhatia, M., et.al. "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.
- [RFC6325] Perlman, R., et.al. "RBridge: Base Protocol Specification", RFC 6325, July 2011.
- [6327bis] Eastlake, D. et.al., "RBridge: Adjacency", draft-eastlake-trill-rfc6327bis, Work in Progress, July 2011.
- [RFC6439] Eastlake, D. et.al., "RBridge: Appointed Forwarder", RFC 6439, November 2011.
- [6326bis] Eastlake, D. et.al., "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-eastlake-isis-rfc6326bis, Work in Progress, December 2011.
- [clearcor] Eastlake, D. et.al., "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-clear-correct, Work in Progress, July 2011.

[IS-IS] ISO/IEC, ''Intermediate System to Intermediate System Routing Information Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)'' ISO/IEC 10589:2002.

8.2. Informative References

[RFC6165] Banerjee, A. and Ward, D. ''Extensions to IS-IS for Layer-2 Systems'', RFC 6165, April 2011.

[RFC4971] Vasseur, JP. et.al ''Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information'', RFC 4971, July 2007.

[TRILLPN] Zhai,H., et.al ''RBridge: Pseudonode Nickname'', draft-hu-trill-pseudonode-nickname, Work in progress, November 2011.

[8021AX] IEEE, ''Link Aggregation'', IEEE Std 802.1AX-2008, November 2008.

[8021Q] IEEE, ''Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks'', IEEE Std 802.1Q-2011, August, 2011

9. Acknowledgments

Authors wish to extend their appreciations towards individuals who volunteered to review and comment on the work presented in this document and provided constructive and critical feedback. Specific acknowledgements are due for Anoop Ghanwani, Ronak Desai, and Varun Shah. Very special Thanks to Donald Eastlake for his careful review and constructive comments.

This document was prepared using 2-Word-v2.0.template.dot.

Appendix A. Change History.

From -01 to -02:

Replaced all references to ''LAG'' with references to Multi-Chassis (MC-LAG) or the like.

Expanded, Security Considerations section.

Other editorial changes.

From -02 to -03

Minor editorial changes

Authors' Addresses

Tissa Senevirathne
Cisco Systems
375 East Tasman Drive,
San Jose, CA 95134

Phone: +1-408-853-2291
Email: tsenevir@cisco.com

Janardhanan Pathangi
Dell/Force10 Networks
Olympia Technology Park,
Guindy Chennai 600 032

Phone: +91 44 4220 8400
Email: Pathangi_Janardhanan@Dell.com

Jon Hudson
Brocade
130 Holger Way
San Jose, CA 95134 USA

Email: jon.hudson@gmail.com

