A Distributed LAG Mechanism for TRILL Enabled Fabrics

Darren Leu
IBM
Santa Clara, CA, USA
dleu@us.ibm.com

Vijoy Pandey
IBM
Santa Clara, CA, USA
vijoy.pandey@us.ibm

***Abstract*—It is desirable to connect external switches or servers to a TRILL campus in a LAG for high availability and for better use of link bandwidth. This paper presents a way, called t-LAG, to allow this to occur via the use of extra RB (called the virtual-RB) for each t-LAG. It makes sense to put multiple t-LAGs together onto a pair of switches, called the t-LAG cluster. All the t-LAGs in a t-LAG cluster can share the same virtual-RB. The use of this virtual-RB for t-LAGs can resolve the load distribution issue for UC traffic. For MC/BC/DLF traffic, a primary link will need to be designated for each t-LAG in order to not receive more than one copy of packets destined to a t-LAG.**

It is recommended to separate the traffic forwarding in a t-LAG cluster into two domains: the TRILL routing domain and the regular L2 switching domain. The traffic handling in these two domains should not be mixed together; i.e., the data switching in the regular L2 domain in a cluster should be handled within the virtual-RB itself. The traffic should not go through the TRILL campus at all if not necessary. The t-LAG ISL for a t-LAG cluster should be used for packet redirection in the regular L2 switching domain when there is a link failure on any t-LAG.

I. INTRODUCTION

Conventional Ethernet networks (known as IEEE 802.1 LANs) use spanning tree protocols (STP) to avoid the loops in a L2 domain; this method imposes a number of new challenges, however, such as inefficient paths, lack of multipath forwarding, etc. [1]. TRILL (Transparent Interconnection of Lots of Links) is an IETF working group intends to resolve these issues with a new set of protocols [2, 3]. TRILL protocols are based on the use of IS-IS [4, 5] in the control plane.

With the use of TRILL protocols [2, 3], regular L2 traffic will be tunneled and passed via routing (a special one, called TRILL routing here) in a TRILL campus. Multi-paths are allowed in a TRILL campus, but not on its boundary – an external switch or server can have only one active link at run time connecting to a TRILL campus for the same VLAN traffic. For high availability, it is desired to have redundant links for external switches or servers to connect to a TRILL campus on more than one RBridge (Routing Bridge). It is also desirable to place these redundant links into a LAG in order to utilize the bandwidth of all the links effectively.

This paper describes a method, called t-LAG (trill-LAG), to allow an external switch or server to connect to a TRILL campus via DMLT (Distributed Multi-Link Trunk).

II. THE PROBLEM STATEMENT

Fig. 1 describes one example of TRILL campus. As noted, multi-paths are allowed inside the TRILL campus, but not on the boundary. If one external switch or server wants to connect to a TRILL campus with more than one link, TRILL protocols will determine the appointed forwarders for the VLANs running on top of the links and, as a result, only one link will be used for data forwarding at run time for one VLAN. For example, Switch 7 in Fig. 1 is connected to both RB4 and RB6 and VLAN 100 is used for data forwarding in both links. If RB6 is chosen as the appointed VLAN-100 forwarder by TRILL, the traffic in between Switch 7 and RB4 will be blocked.

It is the goal to allow external switches or servers to connect to a TRILL campus in a LAG with each link of the LAG connecting to different RBridges. As shown in Fig. 2, we want to connect Switch 7 to both RB4 and RB6 in a LAG.

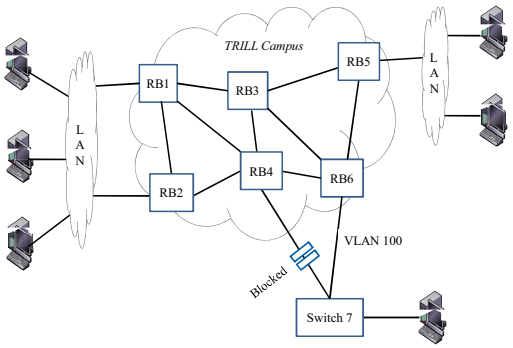

Figure 1 - The VLAN-100 traffic in between Switch 7 and RB4 will be blocked if RB6 is chosen as the appointed VLAN-100 forwarder.

978-0-9836283-2-3

This paper was peer reviewed by subject matter experts for publication in the Proceedings of DC–CaVES 2011

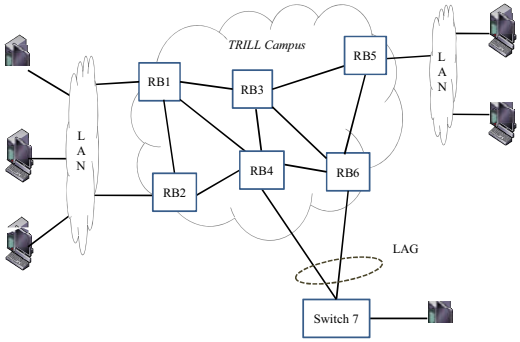

Figure 2 - It is the goal to connect external switches or servers to a TRILL campus in a LAG.

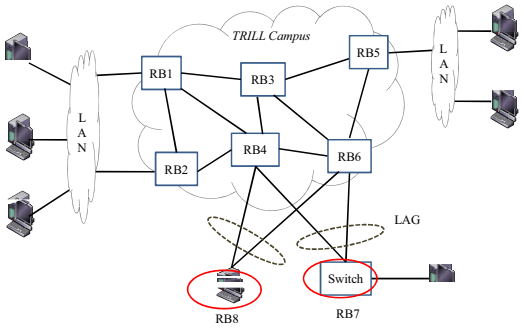

Figure 3 - A t-LAG enabled switch will need to handle traffic for multiple RBridges.

III. THE T-LAG

The t-LAG is designed to resolve the issue above. In the t-LAG design, one extra RBridge should be created and used for each t-LAG configured. This extra RBridge will be called the virtual-RB on this paper. All the virtual-RBs in a TRILL campus should use the same RBridge nickname if they are created for the same t-LAG. All these virtual-RBs for t-LAGs should be involved in the TRILL IS-IS as well as the ESADI communication. A t-LAG enabled RBridge should take care of this communication on behalf of all the virtual-RBs on top of it. SPF computation should also take these virtual-RBs into account, at least for UC traffic.

To support t-LAG, a switch chip should have capability of handling traffic for more than one RBridge. Fig. 3 illustrates one example of such an implementation in which RB4 will need to handle both ingress and egress traffic for RB4 (the RB for the switch itself), RB7 and RB8. RB6 will need to handle the traffic for RB6, RB7 and RB8.

To make it work, for traffic ingress at a t-LAG, the edge RBridges should use the corresponding ingress virtual-RB nickname as the source RB for TRILL encapsulation of the packets. For example, in Fig. 3, the traffic ingress at RB4 may use RB4, RB7 or RB8 as the source RB in the TRILL header depending upon which local port the packet is coming from. Similarly, the traffic ingress at RB6 may use RB6, RB7 or RB8 as the source. In this way, when the packet exits the TRILL campus, the MAC learning performed at egress RBridges will automatically bind the client MAC to the ingress virtual-RB. Once this is done, the load distribution of UC traffic destined to a t-LAG in the campus will be achieved accordingly. Fig. 4 shows one example of such a UC load distribution to a t-LAG. This use of ingress virtual-RB as the source RB in TRILL header, however, may cause problems for MC traffic traversing inside the TRILL campus; this will be discussed later.

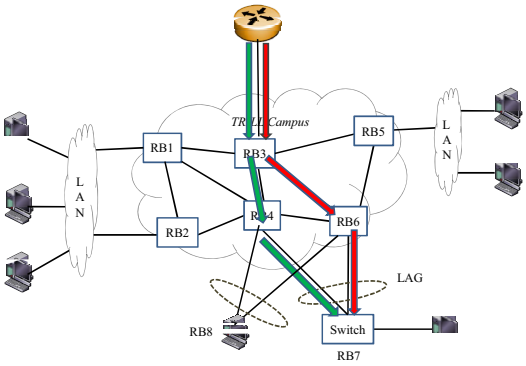

Figure 4 - One example of the load distribution of UC traffic to a t-LAG.

At egress, a t-LAG enabled RBridge should also handle traffic for multiple RBs. Fox example, RB4 in Fig. 3 will need to send traffic out of its local access ports for all the packets destined to RB4, RB7 and RB8; similarly, RB6 will need to handle egress traffic for RB6, RB7 and RB8.

In the control plane, a t-LAG enabled switch will need to handle the TRILL IS-IS communication for all the virtual-RBs on top of it. A CSNP (Complete Sequence Number PDU) will need to be generated automatically by local switch for each virtual-RB on it.

As to the ESADI communication, each t-LAG enabled switch will need to handle all the MAC addresses learnt at its local t-LAG ports.

IV. The t-LAG Cluster

A switch chip today may not handle the TRILL data packets for more than one RBridge. Or, the number of RBridges that can be supported on a chip is usually limited. Besides, the number of distribution trees supported on a chip can also be very limited. Due to these, some adjustments to the above design will be required to adapt to existing hardware.

The idea is to place multiple t-LAGs together onto one pair of RBridges so that these two RBridges will be reserved to serve just for t-LAG purpose. This pair of RBridges will be called the t-LAG cluster on this paper. Note that there can be more than two RBridges in a t-LAG cluster. All the t-LAGs in a t-LAG cluster should share just one virtual-RB and, thus, just one nickname. The total number of RBridges used in the campus will be reduced in this case. Note that a t-LAG cluster can actually use more than one nickname if desired; that is, the assignment of virtual-RB to a t-LAG can be t-LAG based.

Fig. 5 depicts the concept model for connectivity in a t-LAG cluster; in which, the t-LAG cluster consists of two RBridges: RB4 and RB6. RB4 is further divided into two portions: RB4' (the switch RB) and RB4" (the virtual-RB). RB4' will handle the traffic forwarding inside the TRILL campus and RB4" will handle the traffic forwarding outside of the TRILL campus (that is, in the regular L2 switching domain). Similarly, RB6 is further divided into RB6' (the switch RB) and RB6" (the virtual-RB). Furthermore, RB4" and RB6" are combined together into one and share the same virtual-RB, called RB7. RB7 will mainly handle the local switching of L2 traffic for both RB4" and RB6".

If a packet needs to go into the TRILL campus, RB7 will pass the packet to either RB4' or RB6' depending on where the packet is coming from, RB4" or RB6". As noted in Fig. 5, for those traffic that need to pass beyond the TRILL campus, RB4'' is only connected to RB4' and RB6" is only connected to RB6'. The virtual links (RB4' to RB4", and RB6' to RB6") are zero cost and should be handled transparently by the switch chips on RB4 and RB6, respectively.

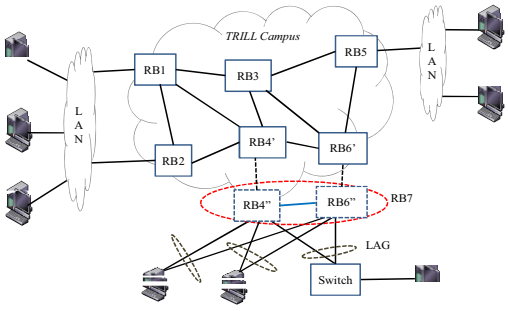

Figure 5 - The concept model for connectivity in a t-LAG cluster.

It is recommended that the handling of all local L2 switching in a virtual-RB (say, RB7) should be handled locally within the RB itself (in this case, RB7). An ISL is required in between RB4 and RB6 (actually, in between RB4" and RB6" in terms of connectivity concept level) for control communication and for failure handling. This ISL is called the t-LAG ISL and is denoted by the blue link in Fig. 5. The ISL will be used for packet redirection whenever there is a link failure on any local t-LAG ports. More on this will be discussed later.

V. Failure Handling

A link in a t-LAG may go down at run time and, due to this, we will need to handle such a failure case to reduce the amount of packet loss to a minimum. There are two ways to help resolve the issue:

1) To adjust at run time the connectivity in between the switch RB (say, RB4' in Fig. 5) and its virtual-RBs (say, RB4" or RB7).
2) To use the t-LAG ISL (in between RB4" and RB6") for packet redirection whenever a failure occurs.

With solution 1, if a t-LAG link drops on a switch (say, one in RB4"),

1) The virtual link in between the switch RB (RB4') and the virtual-RB (RB4"; or, actually, RB7) will be claimed link-down. In this way, after the topology change has been communicated to all other RBridges and a new path has taken effect, the UC traffic routed to RB4 before will now be routed to RB6 for going out via a t-LAG link in RB6".
2) For MC/BC/DLF traffic, the local access ports on edge RBs (RB4" and RB6") will need to be adjusted at run time to allow the traffic be delivered via a healthy link in RB6" for the same t-LAG.

With solution 2:

1) The t-LAG ISL will be used for packet redirection to the peer RB in the same cluster in case a local t-LAG port has a link down.
2) This method can apply to both UC and MC/BC/DLF traffic.

Using solution 1, due to that more than one t-LAG share the same virtual link (say, the one from RB4' to RB4" in Fig. 5), all other healthy t-LAG links on that RB (RB4") will not be used for UC packet delivery once the connectivity in between RB4' and RB4" is claimed link-down. Thus, some bandwidth of healthy t-LAG links will be wasted in this case.

Using solution 2, the t-LAG ISL may get over-loaded if too much traffic needs to pass through it.

As noted, both solutions 1 and 2 should be implemented to better conquer the link failures on t-LAGs. A threshold will need to be implemented and pre-specified in this case so that a t-LAG enabled switch can stop claiming the connectivity in

between the switch RB and the virtual-RB if the number of the local link-down t-LAG ports exceeds the threshold. Note that it will take time for related TRILL IS-IS communication as well as SPF computation to occur and complete before a new topology path can apply once there is a t-LAG link down. Before this occurs, all the traffic delivered to a failed t-LAG link should be redirected as soon as possible via the t-LAG ISL to the peer RB for packet delivery to external switches or servers.

VI. MULTICAST

In TRILL, the multicast traffic is handled differently from that for unicast traffic. A distribution tree will be pre-determined and followed for a specific packet flow of MC/BC/DLF traffic ingress at an RBridge. Usually, all the RBridges in the campus will be visited in all the trees unless VLAN or Multicast pruning has been applied to a tree. It is readily observed that more than one copy of a packet can be delivered to external switches or servers via a t-LAG, if the packet is flooded in the TRILL campus following a tree and all the RBridges will transmit the packet out of its local access ports upon receiving such a packet via TRILL. Actions will be required to prevent this from occurring; that means a primary link for each t-LAG will need to be pre-determined and followed for a specific MC/BC/DLF packet flow. Note that this primary link for a t-LAG is used just for transmission of MC/BC/DLF packets out of a TRILL campus.

There are multiple ways of choosing the primary link for a t-LAG. The selection can be

- System based: The same link in a t-LAG is always used in a campus.
- Distribution tree based: Different distribution trees can use different t-LAG links for MC packet transmission.
- (Distribution tree, VLAN) based: Different t-LAG links can be used for different VLANs in a distribution tree.
- (Distribution tree, VLAN, DMAC) based: Different t-LAG links can be used for different multicast DMAC addresses for the same distribution tree and the same VLAN.

As noted, the selection of the primary link for a t-LAG will need to be adjusted at run time once there is a link up or down in a t-LAG. Some communication for link up and down notification will be required among the RBridges in a t-LAG cluster. Before this adjustment occurs, the t-LAG ISL can be used for packet redirection to avoid the packet drop problem due to that the packet is being sent to a failed t-LAG link.

VII. WHICH RB TO USE AS THE SOURCE RB IN TRILL HEADER?

It is critical to the t-LAG design to bind a client MAC to the ingress virtual-RB for a t-LAG. It would be great if we can use the ingress virtual-RB as the source RB in TRILL encapsulation for a packet when it enters at a t-LAG; the MAC learning performed at egress RBridges will do this binding automatically in this way.

The use of ingress virtual-RB as the source RB in TRILL encapsulation of packets may actually cause problems in MC/BC/DLF distribution in the TRILL campus for some switch chips. Assuming the tree rooted at RB7 in Fig. 6 is used and the link in between RB4' and RB4" is chosen as part of the tree. If a packet gets into the campus via a t-LAG in RB6", the packet may get dropped along the way in the campus (say, by RB3, RB1 or RB4) if the above tree is followed; this is due to that RB7 is used as the source RB on the packet but RB7 is actually on the destination side of the tree.

Instead of using the virtual-RB (RB7) as the source, we should use the switch RB (RB6) as the source RB in TRILL encapsulation in the above case to resolve the issue. Note that both UC and MC traffic should follow the same way for this; otherwise, it will cause MAC learning flapping issue at egress RBridges.

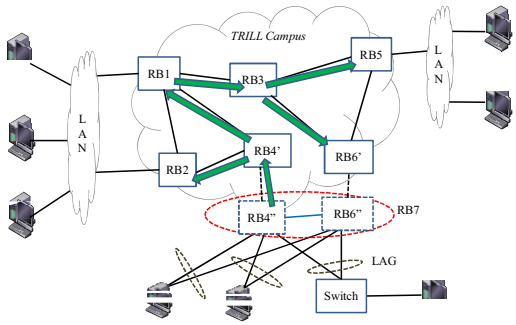

Figure 6 - A MC/BC/DLF packet entering at RB6 may get dropped in the campus following the distribution tree rooted at RB7.

VIII. MAC LEARNING

As mentioned, it is the goal in the t-LAG design to bind the client MAC learnt at a t-LAG to the virtual-RB created for that t-LAG. If the virtual-RB can be used as the source RB in TRILL encapsulation, then this can be automatically done by hardware via the MAC learning performed at egress RBridges. In case we have to use the switch RB as the source in TRILL encapsulation, then a different way for MAC learning will be needed to make this happen.

One way to achieve this is through the software based MAC learning performed on a t-LAG enabled switch. A MAC address learnt at an ingress t-LAG port can be specially manipulated in software to bind to ingress virtual-RB; this newly learnt MAC entry can then be propagated via ESADI to all other RBridges in the TRILL campus for configuration. In this way, the load distribution of UC traffic at any ingress RBridge can then be achieved automatically.

It is also possible to perform the MAC learning via hardware at egress RBridges if the chips there can allow

multiple RBs be mapped onto the same virtual port so that the MAC learning performed on the chips can bind a client MAC to the corresponding ingress virtual-RB.

FDB sync for MAC learnt at t-LAG ports is always required in between the peer RBridges in a t-LAG cluster, especially if the LAG hashing algorithm performed on external switches or servers is SMAC based. This is to avoid unnecessary flooding or dropping of known UC traffic at egress to a t-LAG if the egress RBridge has no related MAC information. The MAC information of the peer RBridge in the same cluster is also needed upon making a decision to redirect traffic to the t-LAG ISL when a local t-LAG link fails.

IX. SOURCE T-LAG PRUNING

Due to that all the RBridges in a TRILL campus will usually be part of a tree most of the time, it is possible that a packet may go back to the t-LAG at which it ingress, through a different link for the same t-LAG at the peer RBridge, of course. Actions such as ACLs will be required on all the t-LAG enabled RBridges to make sure that such a packet be dropped before it goes out.

X. PACKET FLOWS

Packet flows for various scenarios are presented in this section. The scenarios include:

- Case 1: packet flow of UC traffic to a t-LAG via TRILL.
- Case 2: packet flow of MC/BC/DLF traffic from a t-LAG.

Case 1: Packet flow of UC traffic to a t-LAG via TRILL

Fig. 7 illustrates the normal packet flow of some UC traffic via TRILL to a t-LAG. If there is a link failure for that t-LAG in RB4", the packet will be redirected via the t-LAG ISL to the peer RBridge (RB6") in the same cluster for going out through a healthy t-LAG link there, as shown in Fig. 8.

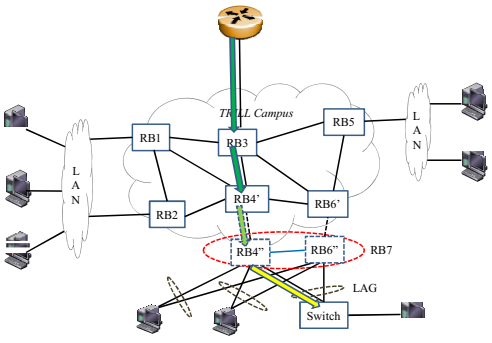

Figure 7 – The normal packet flow of UC traffic to a t-LAG via TRILL.

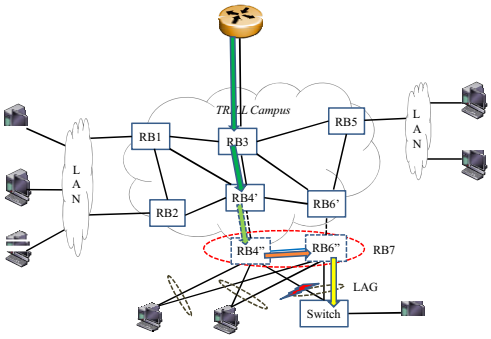

Figure 8 - UC traffic are redirected via the t-LAG ISL to the peer RB for going out in case there is a local t-LAG link failure.

If too many t-LAG links failed in RB4" (the pre-specified threshold is exceeded), the connectivity from RB4' to RB4" will be claimed down; in this case, the UC traffic that routed to RB4 before will take a new route to RB6 for going out, as shown in Fig. 9.

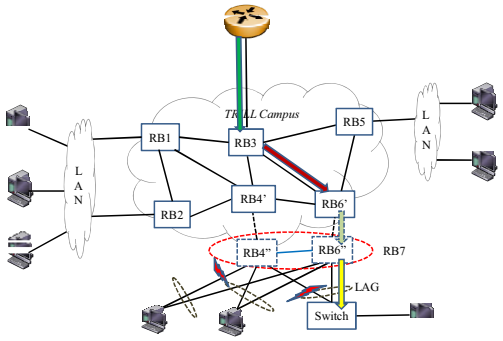

Figure 9 - The UC traffic will take a new route to a t-LAG once the virtual link in between RB4' and RB4" is shut down.

Case 2: Packet flow of MC/BC/DLF traffic from a t-LAG

The normal packet flow for MC/BC/DLF traffic ingress at a t-LAG is shown below in Fig. 10.

If there are local link failures in RB4" as shown in Fig. 11 below, the t-LAG ISL will be used to pass the traffic to the peer RB (RB6") in the same cluster for sending the packets out.

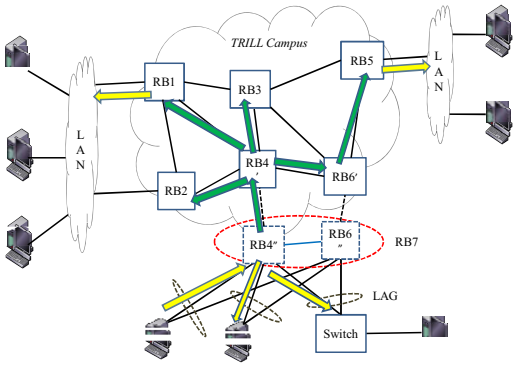

Figure 10 – The normal packet flow for MC/BC/DLF traffic ingress at a t-LAG.

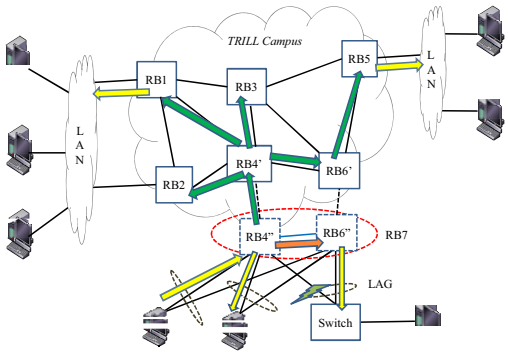

Figure 11 - The t-LAG ISL will be used for packet redirection once a local t-LAG link fails.

XI. SUMMARY

The t-LAG is designed to provide DMLT capability for external switches or servers to connect to a TRILL campus in a LAG; all the links in a t-LAG will be used in an active-to-active way for the same VLAN in this case.

In the t-LAG design, the use of virtual-RB for each t-LAG will make the load distribution of UC traffic to a t-LAG be done naturally. The use of this virtual-RB as the source RB in the TRILL encapsulation will make the MAC learning performed at egress RBridges be done by hardware automatically. The switch chips today, however, may have difficulty to use this method. The switch RB can be used in this case, instead, as the source RB in TRILL encapsulation.

To support t-LAG, ideal switch chips should be capable of handling traffic for multiple RBridges. Existing switch chips today may not have this support in terms of capability or capacity. The use of the t-LAG cluster is to adapt to existing hardware for t-LAG support. All the t-LAGs in a t-LAG need to use just one virtual-RB in this case. A t-LAG cluster can actually consist of more than two RBridges if wanted.

To eliminate the duplicate copies to a t-LAG for MC/BC/DLF traffic, a link in a t-LAG will have to be chosen as the primary link for packet transmission for each specific packet flow. The selection of the primary link for a t-LAG can be: system based, or a combination of distribution tree, VLAN, and DMAC based. Actions are required at egress RBridges to make sure a packet won't go back to its originating t-LAG.

The traffic handling in a t-LAG cluster should be separated into two domains: one for the traffic routing within the TRILL campus and the other for the traffic switching in the regular L2 domain. It is recommended to totally separate the traffic handling in these two domains in a t-LAG cluster. The t-LAG ISL will be required in a t-LAG cluster between peer RBs to handle the traffic redirection once there is any local link failure on t-LAGs. The traffic redirection via the t-LAG ISL is required until a new route or distribution for affected traffic be determined and applied – this will take time to occur.

REFERENCES

[1] J. Touch and R. Perlman, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement", RFC 5556, May 2009.

[2] R. Perlman, D. Eastlake 3rd, D. Dutt, S. Gai, A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[3] R. Perlman, D. Eastlake, Y. Li, A. Banerjee, H. Fangwei, "RBridges: Appointed Forwarders", <draft-ietf-trill-rbridge-af-04.txt>, Internet-Draft, expires: January 6, 2012.

[4] D. Eastlake, A. Banerjee, D. Dutt, R. Perlman, A. Ghanwani, "TRILL Use of IS-IS", RFC 6326, July 2011.

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