TRILL using Pseudo-Wire Emulation (PWE) Encapsulation

<draft-bryant-perlman-trill-pwe-encap-00.txt>

Stewart Bryant <stbryant@cisco.com>
Radia Perlman <Radia.Perlman@sun.com>
Alia Atlas <akatlas@alum.mit.edu>
Don Fedyk <dwfedyk@nortelnetworks.com>

Motivation

It is assumed that TRILL encapsulation requires

- TTL
- RBridge ID, (could be the ingress or the egress)

Four encapsulation mechanism that TRILL could use:

1. It could design its own encapsulation from scratch.
2. It could use an Ethernet based encapsulation.
3. It could use an IP based encapsulation.
4. It could use an MPLS based encapsulation.

Forwarding Compatibility

- Adding/removing encap, & forwarding is one of the most time critical operation in any networking equipment.
- Forwarding usually requires hardware support.
- New network encapsulation type that needs new hardware is
  - Expensive to design and deploy
  - Significant time and risk impact on the market acceptance of a new network architecture.

Avoid a new TRILL specific encapsulation, if possible.

Ethernet Encap

- The nesting of 802.x tags is a well understood technology and suitable hardware is widely deployed.
- Absence of a TTL field means traditional convergence mechanisms will create loops.
- Existing controlled convergence techniques may resolve this issue – but are not currently widely understood.

IP Encapsulation

- IP encapsulation issues (see Section 5.5 in [RBRIDGE]).
  - Encapsulation overhead
  - Complexity of providing L2 services within the L3 subnet
  - Potential fragmentation and reassembly work

Such issues mean IP is not an acceptable encapsulation format.

RFC3032 Encapsulation

- The proposal is **NOT** to use MPLS
- The proposal is the reuse of the MPLS Label Stack Entry mechanism that is widely supported in existing hardware.
- Provides
  - 19 bit Source or Destination Address
  - 1 bit Address type (unicast/other) indicator
  - TTL
  - QoS indicator
- Supports approx 500,000 addresses
- Stacking support *might* be useful for extensibility

Ethernet over RFC3032

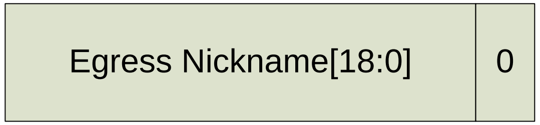
```
0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Egress Nickname                 |0| Exp |S|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|              Received Ethernet Frame                        ///
///                                                             |
|                                                               |
+---------------------------------------------------------------+

                     Exp:    Indicates Priority
                     S:      Bottom of Stack, 1 bit
                     TTL:    Time to Live, 8 bits
```

PWE3 uses this encapsulation.

Unicast Traffic

When an RBridge X learns a new egress nickname, form the 20-bit label field.

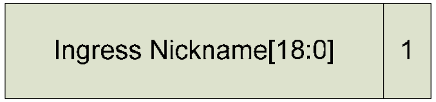

Egress Nickname[18:0]
0

Label Field

- Create a global in-segment for traffic received with the label field specified with the nickname + unicast marker.
- For each interface on the shortest path tree from RBridge X to the Rbridge indicated by A, create an out-segment that swaps to that label field.
- Connect the in-segment to those out-segments with load-balancing specified; only one out-segment will be used for a particular frame.
- For forwarding plane compatibility, QoS can be obtained from the EXP field.

Multicast and Broadcast

- Ingress RBridge nickname indicates the spanning tree to should be used.
- Just like unicast, label field is formed from nickname and unicast/other marker.



Label Field

- When an RBridge learns of a new ingress RBridge nickname, an ILM entry corresponding to the label is created.
- An out-segment is created for each interface that is in the SPT rooted at the ingress RBridge.
- The in-segment is connected to the created out-segments with multicasting specified; subject to filtering, each frame will be sent out each out-segment.
- Except for the egress filtering, the above forwarding behavior is already defined and available in some forwarding planes.

Egress Filtering

- Transmission of a multicast or broadcast packet may be subject to egress filtering.
- For instance, if a broadcast frame is tagged with a VLAN and an interface is marked as not being part of the VLAN or connected to any RBridges or bridges on the VLAN, then the frame may be dropped instead of sent.
- Similarly, an RBridge could decide to filter a multicast frame instead of sending it, if the interface were known to no be part of the multicast tree.

Dynamic Nickname Selection

- Each RBridge has a unique 6-byte MAC system ID, which it uses as its IS-IS ID.
- An RBridge learns the topology (and advertised nicknames) before picking its own 19 bit nickname.
- Each RBridge is also responsible for ensuring that its nickname is unique.
    - If R1 chooses nickname x, and R1 discovers, through receipt of R2's LSP, that R2 has also chosen x, then the RBridge with the lower system ID keeps the nickname, and the other one must choose a new nickname.
- Nicknames may change on network merge but R1 and R2 (above) are aware as soon as they see new LSPs; the same tie-breaking mechanism as above can be used.

Selection Algorithms

- Choice of selection algorithm a local matter
- Example methods:
  - Randomly from the set of assigned nicknames
  - hash algorithm
  - Etc. (see draft)

Summary

- RFC3032 encapsulation provides:
  - TTL, QoS, rbridge nicknames
  - No forwarding path changes required
  - Minor forwarding path changes needed for intelligent load-balancing
  - Minor forwarding path changes desirable for efficient multicast/broadcast filtering
  - Minimal overhead