Ingress Replication P-Tunnels in MVPN

- Ingress Replication has always been one of the P-tunnel technologies supported by MVPN
- But there's a lot of confusing text in the documents
  - Sometimes an IR tunnel is discussed as if it were just a set of unicast tunnels
  - But there are places in the spec where one is told to:
    - advertise the tunnel on which you will send a given flow
    - discard packets from the wrong PE (how do you know the ingress PE of a unicast tunnel, if it's an LDP-created LSP)
    - discard packets that come from an unexpected tunnel (extranet)
    - change the upstream multicast hop for a given tunnel (i.e., prune yourself from a given tunnel and rejoin it at a different place)
- This text is about some kind of P2MP tunnel, not about unicast tunnels
- There seems to be some concept of IR tunnel in which an IR tunnel consists of a set unicast tunnels, but is not itself a unicast tunnel

IR Tunnels and the PMSI Tunnel Attribute

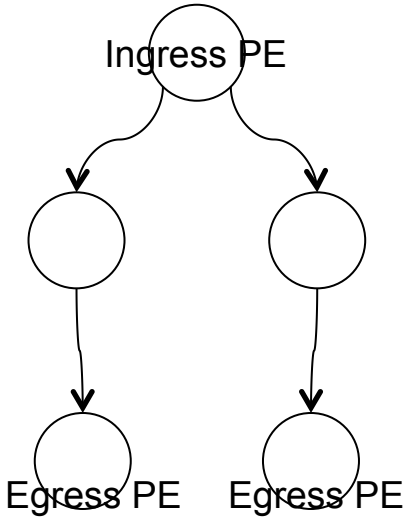
- PMSI Tunnel Attribute (PTA) has:
  - Tunnel type
  - Tunnel identifier
  - MPLS label
- In I/S-PMSI A-D route, if type is IR, identifier field is unused!
- In Leaf A-D route, if type is IR, identifier is the IP address of the originator of the route
- Isn't there an identifier for the IR tunnel itself?
  - If so, where is it?
  - If not, what does it mean to join and/or leave an IR tunnel, or to change one's UMH for a given IR tunnel?
- The MPLS label in the PTA of a Leaf A-D route is used, but are there any requirements on the label allocation policy? Can the PTAs of different Leaf A-D routes use the same label?

Purpose of the IR Draft

- When implementing/deploying IR capability, we discovered quite a few questions whose answers were not obvious
- draft-rosen-l3vpn-ir-00.txt attempts to clear up the issues around IR tunnels by:
  - establishing a clear conceptual model for IR tunnels
  - explaining how an IR tunnel is identified
  - explaining how to join/leave an identified IR tunnel
  - how to apply the discard from the wrong PE or wrong tunnel policies to IR tunnels
  - setting out the requirements on MPLS label allocation
  - explaining how to switch from one IR tunnel to another in “make before break” fashion
  - explaining how to change your UMH within a given IR tunnel, again in “make before break” fashion.

What is an IR Tunnel?

- Think of an IR tunnel as a P2MP tree, where traffic from a given parent node to a given one of its child nodes is carried through a unicast tunnnel
  - If non-segmented tunnels are used, the root node of an IR tunnel is an ingress PE, and its children are egress PEs
  - If segmented tunnels are used, one can think of the IR tunnel as a multi-level P2MP tree, with ABRs/ASBRs as intermediate nodes
- Each node on an IR P2MP tree maintains multicast state for that tree
- Each edge is a unicast tunnel, consisting of a sequence of routers that do not maintain multicast state for this tree
- The unicast tunnels may carry packets of multiple IR tunnels, along with real unicast packets



IR Tunnel Setup Protocol

- IR is unique in being the only P2MP tunnel type that doesn't come with a setup protocol of its own
- All setup is done using MVPN BGP A-D routes
  - Advertise via I/S-PMSI A-D route
  - To join a tree, choose a parent node, create a Leaf A-D route identifying the tree, and "target" it to the parent node
  - Leaf A-D route is "targeted" to a given parent node by attaching an IP-address-specific RT identifying the parent node
- But to join a tree, you have to identify it. Unlike other tunnel types, the PTA contains no identifier of the tree. Where's the identifier?
- The identifier of an IR tunnel is the NLRI of the I/S-PMSI A-D route that announces it
- The Leaf A-D route carries that identifier in its own NLRI (as the "route key"), which is how it specifies the tree it is trying to join.

What goes in the Leaf A-D Route PTA?

- When Leaf A-D route is sent from child to parent, RT identifies parent, child identified in both NLRI and PTA “tunnel id” field
- Not much information provided about the unicast tunnel between parent and child
  - only child IP address provided
  - unicast tunnel type must be known *a priori*
- Child provides MPLS label (downstream-assigned) that parent uses when transmitting through the IR tunnel to the child
  - MPLS label field of Leaf A-D route PTA
  - On data packets, label is carried inside a unicast encapsulation (which is likely to itself be MPLS, possibly with implicit null)
- Interesting factoid: can’t use S-PMSI A-D route to assign two C-flows to the same IR tunnel
  - MPLS Label field in PTA of I/S-PMSI A-D route has no use

MPLS Label Allocation Policy (1)

- Every IR tunnel has a “root” and a “root RD”
  - Root is either ingress PE or (or for IR tunnels advertised in Inter-AS I-PMSI A-D routes) ingress AS
  - Can be inferred from tunnel identifier (NLRI); details in draft
- Egress PE policies:
  - Never assign same label to IR tunnels that have different roots
    - Otherwise “discard from wrong PE” policy cannot be applied
  - If changing parent nodes on a given tree, change the label also
    - During the transient, one may receive duplicate packets, as old and new parents may both be transmitting
    - Need to use different labels to ensure that one of the duplicates is discarded

MPLS Label Allocation Policy (2)

- Acceptable Egress PE policy for non-extranet:
  - Label unique per <root, parent, egress VRF>
  - Allows “discard from wrong PE” policy to be applied
  - Prevents duplicates during transient changes
  - Allows dispatch to proper VRF context
- Acceptable Egress PE policy for extranet:
  - Label unique for each <root, RD of root, parent>
  - Need uniqueness per ingress VRF, to apply “discard from wrong P-tunnel” policy that is needed for extranet
  - Allows dispatch to multiple VRFs
  - Prevents duplicates during transient changes

MPLS Label Allocation Policy (3)

- Intermediate nodes receive Leaf A-D routes from child nodes
- Two Leaf A-D routes (from two child nodes) with same IR tunnel identifier (route key in NLRI) result in only one Leaf A-D route with that route key being sent upstream
- Safe policy: assign unique label per route key
- But strictly speaking:
  - If multiple IR trees all have only one child node, and it's the same for all, and that child node has assigned the same label to all those trees, intermediate node can also assign a single label to all those trees (as long as this condition continues to hold)

Make before Break (1)

- *Make before break* is desirable when:
  - Changing the IR tree on which a given C-flow is to be received
  - Changing one's parent node on a given IR tree
- To change parent node on given IR tree, change the RT on the Leaf A-D route used to join that tree
- Effect*: simultaneously* (and immediately) prunes from the old parent and joins via the new parent
- But to do make before break, we want to:
  - keep receiving traffic from the old for awhile
  - join the new, but discard traffic from the new for a while
  - start accepting traffic from the new, but discard from the old
  - prune from the old
- Can't do this with the control plane, because a single BGP path attribute change causes both the "join the new' and the "prune from the old"

Make before Break (2)

- Make before break must be done with data plane timers
- Parent node actions:
  - When a child node prunes itself from an IR tree, old parent node keeps transmitting to it on that tree, for a period of time
  - When a child node joins a tree via a particular parent, new parent begins transmitting immediately
- Child node actions:
  - When joining a tree via a particular parent, and already joined via a different parent, for a period of time discard from new parent but accept from old parent
  - After a period of time, discard from old parent but accept from new parent
  - Note that this requires different labels to be advertised to the two parents
  - Note also that there is no way to send a Leaf A-D route to both parents at the same time, as each Leaf A-D route has only one PTA and thus assigns only one label