The Evolving MPLS Forwarding Model

Stewart Bryant

sb@stewartbryant.com

Why the PALS+ Meeting

- There have been a recent cluster of proposed MPLS changes
  - A number of proposals to add metadata after the stack
  - A proposal to add metadata in the stack
  - A number of proposal to create new Extended Special Purpose Labels (eSPL) (SPLs used to be called Reserved Labels)
  - A proposal to repurpose some (e)SPLS bits

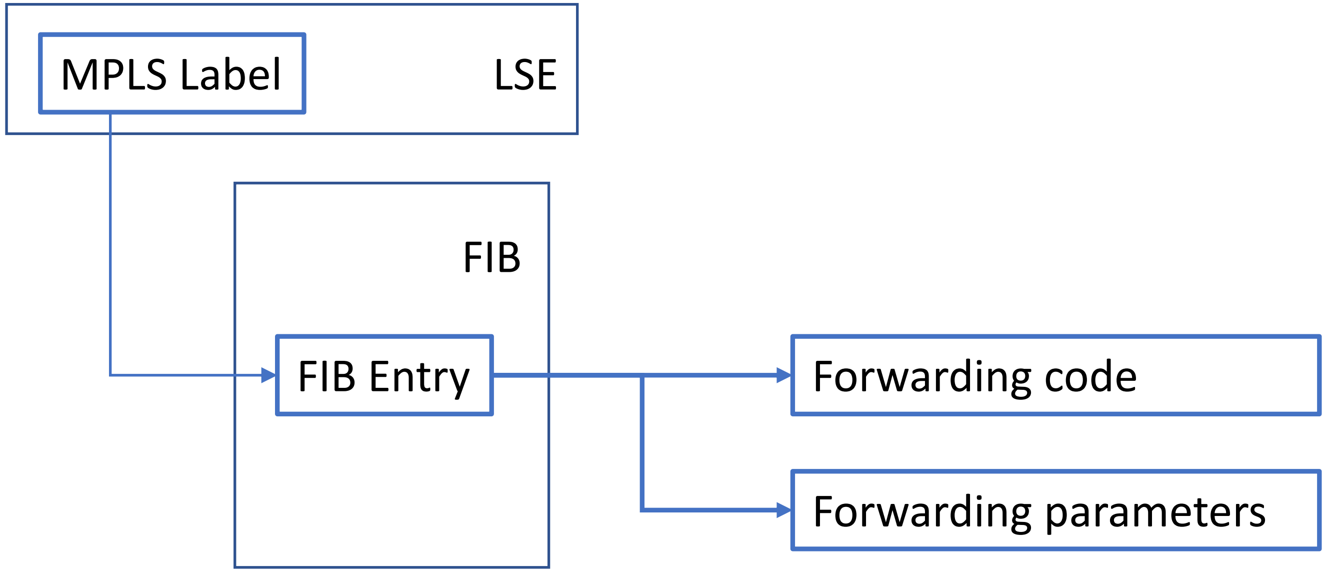
Why PALS

- PALS was the first WG to deploy metadata below Bottom of Stack. (Pseudowire Control Word)
- DetNet copied this approach (DetNet Control Word)
- Some new proposals would require more than one set of metadata below the Bottom of Stack.
- PALS needs to make sure that the addition of other metadata does break or change its protocols, OR, if it does, we know:
  - How the behaviour changes
  - What the consequences, costs and mitigations are.

The Architectural View

- None of the proposed changes are large or significant in themselves
- However, in combination:
  - Results in complexity
  - Change the MPLS forwarding model
  - May fundamentally limit MPLS development
- So we called the interested parties together to:
  - Understand the needs that underpin the proposals
  - Understand the proposals
  - Develop a way forward that deliberately changes the MPLS architecture to a model that meets our needs, rather that discover that we have boxed ourselves into a corner as a result of a set of uncoordinated changes.

The Original Model

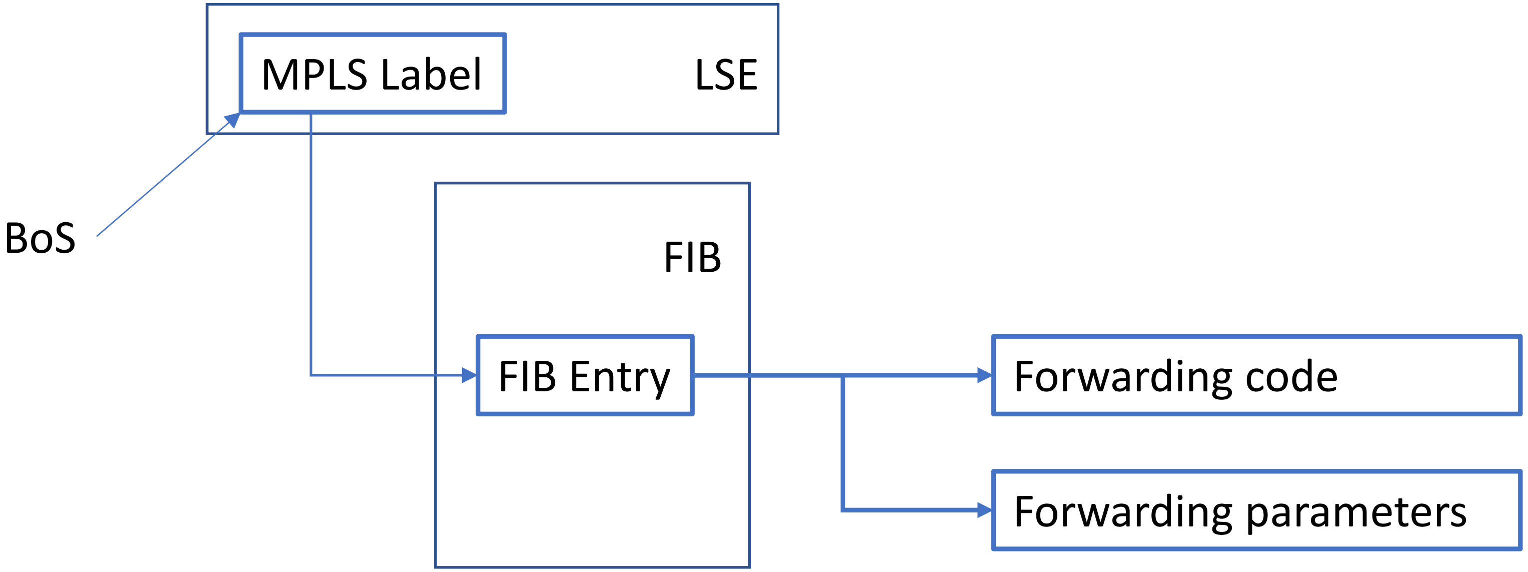


The model is simple and general

- Only the top label is processed
- New forwarding code can be added and the label that activates it is included in the packet.
- The mapping between the label and the forwarding code and any forwarding parameters is provided by the control plane.

LSE = Label Stack Entry (what many people call a label)
FIB = Forwarding Information (date)Base

VPN and PW



VPN: Fwd code processes IP payload in the context provided by the VRF label.

PW: Fwd code processes payload in the context provided by the PW label. Parameters say if there is a CW.

CW: A block of data below the Bottom of Stack that provided additional per packet parameters.

Original Special Purpose Labels

- Single label at Bottom of Stack.
- Only processed at disposition, i.e. when all previous labels popped
- Exp Null, RA, GAL

ECMP – The End of the ”Pure” Model

- Equal Cost Multi-path (ECMP) is needed to balance traffic over a set of available next hops to minimize congestion by spreading the traffic over the available paths.
- Approaches
  - Hash the label stack
    - No longer a pure Top of Stack forwarding model, but the processing is optional
  - Add an entropy label at BoS, remove as part of processing BoS Label
  - Walk the stack, holistically test if payload is IP, hash the IP five tuple
    - … and make mistakes, so first nibble ECMP avoidance was added.
  - Walk the stack to find the ELI (from the original SPL set), if found load balance on the EL that follows.

Extended Special Purpose Labels

- With half the SPLs used and SPLs becoming popular eSPLs were created.
- Label pairs <Extn Label = 15> <Label = 0..255>
- Problems:
  - Two labels to push increasing size of stack to parse to find BoS
  - Two labels = two tests in forwarder.
  - As number of labels used increases eventually need new h/w to do lookup
- E(SPL)s seen by some as a means of avoiding control plane operations at the expense of forwarding efficiency, the antithesis of the original MPLS design.

Impact of Multiple SPLs

- Stack space
    - Particularly a problem for EL and ESPL which take two LSEs
    - Stack space is very limited on some edge routers - as short as five labels
        - FRR, Delivery, VPN, ELI+EL – no more room
    - All routers have a maximum stack depth view
        - @ two labels a time makes finding BOS harder
        - Complex processing when pushing additional label (Tunnel of FRR) and making sure that existing SPLs are in view downstream.
- Complexity of correct processing order to be considered

Data Past the End of Stack

- Control word to provide additional per packet processing parameters to disposition router.
  - PW Control Word
  - DetNet Control Word.
- OAM instruction/Data (the G-ACH mechanism)
  - Mutually exclusive with user data
  - No more than one ACH per packet.

Changes

- There were two proposals in the works when we called the meeting:
  - Fragmentation
  - iOAM
- Both propose to add additional data past BoS
- Both at some stage proposed to indicate presence by ESPL
- Both propose to run on packets carrying user data
- One (iOAM) requires action per hop and at egress.
- Both could be on the same packet (together with CW)
- There are now other proposals in gestation.

Questions

1. Should we support the use of SPL embedded in the stack to trigger per hop behavior in MPLS or should we require all per hop behavior to be triggered by the Top of Stack label?
2. Should we support SPL at Top of Stack to trigger per hop behaviour?
3. When should we support the use of SPL to trigger BoS behaviour in MPLS and when should we require use of a regular label?
4. Should we support multiple separate ACH below BoS? If so, should they be indicated individually or as a single entity (I.e. One MPLS label or several stacked MPLS labels).

Backup - The proposals (probably incomplete)

- draft-andersson-mpls-eh-label-stack-operations
- draft-andersson-mpls-eh-architecture
- draft-cheng-mpls-inband-pm-encapsulation
- draft-gandhi-mpls-ioam-sr
- draft-kompella-mpls-mspl4fa
- draft-kompella-mpls-nffrr
- draft-li-mpls-enhanced-vpn-vtn-id
- draft-song-mpls-eh-indicator
- draft-song-mpls-extension-header
- draft-xu-mpls-payload-protocol-identifier