draft-villamizar-mpls-tp-multipath-01

Curtis Villamizar (Infinera)

Briefly this internet-draft:

1. Better documents common multipath practices.

2. Documents scaling and capacity efficiency motivations

3. Proposes to better accommodate MPLS-TP with simple extensions.

What is Multipath?

Multipath includes:

1. Equal Cost Multi-Path (ECMP)

2. Link Aggregation (LAG)

3. Link Bundling using all ones component link.

Note: ECMP is widely used but not applicable to MPLS-TP.

Apparent (but solvable) Dilemma (see Section 1)

- MPLS-TP OAM as defined does not work on multipath.
  - CCV fate sharing is affected.
  - LM accuracy is affected.
  - There are other impacts.
- Core networks make extensive use of ECMP and LAG.
  - Providers claim to have 30x 10GbE and 40x 10GbE links.
  - Link bundling placing each LSP on only one component is inefficient and scales poorly (see next slides)

Multipath Related Requirements (see Section 3.1)

R#1 to R#5 (paraphrased for brevity) are related to Multipath.

R1 Multipath MUST exceed largest possible component (100G today)

R2 Multipath SHOULD carry LSP bigger than largest possible component.

R3 MPLS hierarchy SHOULD NOT be constrained by multipath limitations.

R4 Multipath SHOULD carry LSP bigger than largest single packet processing element (today 100G, 400G not yet)

R5 Load split SHOULD use capacity very efficiently.

MPLS-TP Related Requirements (see Section 3.2)

R#6 to R#7 are related to MPLS-TP. R#8 to R#11 are related to MPLS-TP scaling. Paraphrased for brevity.

R6 Traffic within a MPLS-TP MUST NOT be reordered unless explicitly allowed. [by configuration or signaling]

R7 Traffic within a MPLS-TP MUST NOT be reordered if full OAM capability is required.

R8 It must be possible to aggregate MPLS-TP LSP into bigger LSP.

R9 MPLS-TP OAM and payload must fate share even when aggregated.

R10 MPLS-TP constraints which limit use of MPLS hierarchy SHOULD be optional.

R11 LSP carrying MPLS-TP LSP SHOULD be able to exceed the size of a single multipath component link.

Scaling Issues (see Section 3.3)

The remainder of section 3 gives scaling related reasons for some of the requirements. This is clearly marked as "supporting discussion". Briefly:

1. Not using multipath makes these scaling issues worse:
   - ILM size
   - CSPF scaling
2. Aggregating traffic using topological hierarchy improves scalability but creates very large LSP.
3. LSP capacity of over 100G exists today (carried by N x 10G LAG). Many provider links are well over 100G.

Scaling Issue: ILM size

Consider an N node graph with cutset in middle.

Roughly half the nodes are on either side of the cutset.

A full mesh of LSPs requires $1/4N^2$ LSP across cutset.

A cutset may be as few as 2-3 right of ways.

If path protection is used double the number of LSP.

For N = 2,000 the cutset saturates the 20 bit label space for a cutset of two.

A classic example is US E-W paths west of Mississippi River, where many core networks have a cutset of two or three.

Scaling Issue: CSPF scaling

Consider N nodes in an IGP area and MPLS core full mesh.

Each node is ingress to N-1 LSP (1/2 signaling for bidir LSP).

Each CSPF computation is order $Nlog(N)$ (where $L = kN$, k is node degree).

If a large number of LSP go down due to a fault, order $N^2log(N)$ computation is required.

Note: Dynamic routing is always provided in IP/MPLS networks to support restoration, and therefore recovery from multiple faults. Speed of restoration is important.

Scaling Issue: Summary

Both the ILM and computation (CSPF) issues are documented in RFC 5439 "An Analysis of Scaling Issues in MPLS-TE Core Networks".

Solutions in RFC 5439 are:

1. use LDP (use MP2P, incompatible with MPLS-TP)
2. use RSVP-TE hierarchy (reduces signaling and state)

Use of RSVP-TE hierarchy creates a much smaller number of much larger LSP (over 100G) carried over very large multipath links (300-400G today, multiple Tb/s in a few years). Control plane scaling therefore requires multipath.

Current Practices (see Section 4)

- RFC 4928, Section 2, "Current ECMP Practices", is cited.
- Flow identification (actually groups of flows) is based on:
  1. hash over IP src/dst pairs, or
  2. hash over label stack.
- Simple vs adaptive multipath
  - simple multipath just does hash and modulo or equiv
    - no feedback on accuracy of balance is used
  - adaptive multipath finely tunes load split
    - feedback is internal to NE
    - no change in standards is required
    - supported by some product since very late 1990s
- Parallel link or parallel VIF may be used.
- Additional best practices are described Section 4.2.

Addressing Requirements (see Section 5)

MPLS-TP and multipath MUST coexist somehow or MPLS-TP will not be applicable to core network use. Three options given:

1. MPLS-TP client layer over MPLS server layer.

   [small changes to current practices provides best solution.]

2. MPLS client layer over MPLS-TP server layer with MPLS client layer using multipath as-is.

   [scales poorly, uses bandwidth inefficiently]

3. Relax MPLS-TP OAM requirements.

   [This is unacceptable to some providers.]

Pros and Cons of Approaches

This is summarized in Section 5.1.2 with further detail in Section 5.2.

	Advantage	Disadvantage
MPLS server layer	Supports fully compliant MPLS-TP as client. Efficient use of capacity.	Requires data plane change.
MPLS-TP server layer	Some transport vendors prefer this.	Inefficient. Poor scaling, may be unusable in large core.
Relax MPLS-TP OAM	Allows MPLS-TP over multipath.	OAM functionality is impaired. Unacceptable to some providers.

Data plane changes

- Existing practice:
  - IP src/dst hash
  - look past labels for IP
  - MPLS label hash
- New practices:
  - Disable hash for some LSP
  - Disable hash on IP payload for some LSP
  - Limit hash depth from top of label stack for some LSP
  - skip over reserved labels when doing hash
  - signaling indicates desired behaviour
- Why this works:
  - No hash done on top level MPLS-TP LSP (disabled)
  - No hash done on payload for MPLS-TP LSP (IP hash disabled)
  - No hash done on PW label for MPLS-TP LSP (depth limited)
  - No hash done on MPLS-TP LSP within MPLS LSP (depth limited)
  - OAM and payload fate-sharing (GAL does not affect hash).

Control plane changes

A TLV is needed in link or FA advertisements to indicate MPLS-TP capability on transmit side of a link bundle.

MPLS-TP LSP must be identified as MPLS-TP in signaling (indicate hash based load split cannot be done on this LSP).

LSP signaling must indicate if IP or PW without CW is being carried (suppress IP based hash to support PW payload).

MPLS LSP containing MPLS-TP LSP must signal hash depth limit to prevent hash based load split of MPLS-TP.

These are very modest changes (a few bytes per link or LSP)

Recommendations in draft

Forwarding and signaling changes (previous slide) are recommended.

Allow labels below GAL (ignore BOS when GAL is encountered and look for ACH after the BOS).

Option to relax OAM is recommended to accommodate old hardware that will always hash the whole stack plus IP.

Some changes to other documents recommended to minimize impairment to OAM if it is required to relax OAM in some parts of a network.

Request to WG

Please accept this draft as a WG item.

This would be informational (requirements and framework).

Data plane practices remains informational but could be split out into a separate informational or BCP document.

Protocol extensions would be separate internet-drafts.