Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 18, 2009

N. So
A. Malis
D. McDysan
Verizon
L. Yong
Huawei USA
F. Jounay
France Telecom

February 14, 2009

Framework and Requirements for Composite Transport Group (CTG)
draft-so-yong-mpls-ctg-framework-requirement-01

Status of this Memo

Copyright Notice

Abstract

This document states a traffic distribution problem in today's IP/ MPLS network when multiple physical or logical links are configured between two routers. The document presents a Composite Transport Group framework as TE transport methodology over composite link for the problems and specifies a set of requirements for Composite Transport Group(CTG).

Table of Contents

1. Introduction

IP/MPLS network traffic growth forces carriers to deploy multiple parallel physical/logical links between two routers. The network is also expected to carry some flows at rates that can approach capacity of any single link, and some flows to be very small compared to a single link capacity. There is not an existing technology today that allows carriers to efficiently utilize all parallel transport resources in a complex IP/MPLS network environment. Composite Transport Group (CTG) provides the local traffic engineering management/transport over multiple parallel links that solves this problem in MPLS networks.

The primary function of Composite Transport Group is to efficiently transport aggregated traffic flows over multiple parallel links. CTG can take the flow TE information into account when distributing the flows over individual links to gain local traffic engineering management and link failure protection. Because all links have the same ingress and egress point, CTG does not need to perform route computation and forwarding based on the traffic unit end point information, which allows for a unique local transport traffic engineering scheme. CTG can transport both TE flows and non TE flows. It maps the flows to CTG connections that have assigned TE information either based on flow TE information or auto bandwidth measurement on the connections. CTG distribution function uses CTG connection TE information in the component link selection that CTG connections traverse over.

This document contains the problem statements and the framework and a set of requirements for TE transport methodology over composite link. The necessity for protocol extensions to provide solutions is for future study.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2.1. Acronyms

BW: BandWidth

CTG: Composite Transport Group

ECMP: Equal Cost Multi-Path

FRR: Fast Re-Route

LAG: Link Aggregation Group

LDP: Label Distributed Protocol

LR: Logical Router

LSP: Label Switched Path

MPLS: Multi-Protocol Label Switching

OAM: Operation, Administration, and Maintenance

PDU: Protocol Data Unit

PE: Provider Edge device

RSVP: ReSource reserVation Protocol

RTD: Real Time Delay

TE: Traffic engineering

VRF: Virtual Routing & Forwarding

2.2. Terminologies

Composite Link: a group of component links that acts as single routable interface

Component Link: physical link (e.g. Lambda, Ethernet PHY, etc) or logical links (e.g. LSP, etc)

Composite Transport Group (CTG): traffic engineered transport function entity over composite link

CTG connection: a connection used for data plane

3. Problem Statements

Two applications are described here that encounter problems when multiple parallel links are deployed between two routers in today's IP/MPLS networks.

3.1. Incomplete/Inefficient Utilization

An MPLS-TE network is deployed to carry traffic on RSVP-TE LSPs, i.e. traffic engineered flows. When traffic volume exceeds the capacity of a single physical link, multiple physical links are deployed between two routers as a single backbone trunk. How to assign LSP traffic over multiple links and maintain this backbone trunk as a higher capacity and higher availability trunk than a single physical link becomes an extremely difficult task for carriers today. Three

methods that are available today are described here.

1. A hashing method is a common practice for traffic distribution over multiple paths. Equal Cost Multi-Path (ECMP) for IP services and IEEE-defined Link Aggregation Group (LAG) for Ethernet traffic are two of the widely deployed hashing based technologies. However, two common occurrences in carrier networks often prevent hashing being used efficiently. First, for MPLS networks carrying mostly Virtual Private Network (VPN) traffic, the incoming traffic are usually highly encrypted, so that hashing depth is severely limited. Second, the traffic in an MPLS-TE network typically contain a certain number of traffic flows that have vast differences in the bandwidth requirements. Furthermore, the links may be of different speeds. In those cases hashing can cause some links to be congested while others are partially filled because hashing can only distinguish the flows, not the flow rates. A TE based solution better applies for these cases. IETF has always had two technology tracks for traffic distribution: TE-based and non-TE based. A TE based solution provides a natural compliment to non-TE based hashing methods.

2. Assigning individual LSPs to each link through constrained routing. A planning tool can track the utilization of each link and assignment of LSPs to the links. To gain high availability, FRR [RFC4090] is used to create a bypass tunnel on a link to protect traffic on another link or to create a detour LSP to protect another LSP. If reserving BW for the bypass tunnels or the detour LSPs, the network will reserve a large amount of capacity for failure recovery, which reduces the capacity to carry other traffic. If not reserving BW for the bypass tunnels and the detour LSPs, the planning tool can not assign LSPs properly to avoid the congestion during link failure when there

are more than two parallel links. This is because during the link failure, the impacted traffic is simply put on a bypass tunnel or detour LSPs which does not have enough reserved bandwidth to carry the extra traffic during the failure recovery phase.

3. Facility protection, also called 1:1 protection. Dedicate one link to protect another link. Only assign traffic to one link in

the normal condition. When the working link fails, switch traffic to the protection link. This requires 50% capacity for failure recovery. This works when there are only two links. Under the multiple parallel link condition, this causes inefficient use of network capacity because there is no protection capacity sharing. In addition, due to traffic burstiness, having one link fully loaded and another link idle increases transport latency and packet loss, which lowers the link performance quality for transport.

None of these methods satisfies carrier requirement either because of poor link utilization or poor performance. This forces carriers to go with the solution of deploying single higher capacity link. However, a higher capacity link can be expensive as compared with parallel low capacity links of equivalent aggregate capacity; a high capacity link can not be deployed in some circumstances due to physical impairments; or the highest capacity link may not large enough for some carriers.

An LDP network can encounter the same issue as an MPLS-TE enabled network when multiple parallel links are deployed as a backbone trunk. An LDP network can have large variance in flow rates where, for example, the small flows may be carrying stock tickers at a few kbps per flow while the large flows can be near 10 Gbps per flow carrying machine to machine and server to server traffic from individual customers. Those large traffic flows often cannot be broken into micro flows. Therefore, hashing would not work well for the networks carrying such flows. Without per-flow TE information, this type of network has even more difficulty to use multiple parallel links and keep high link utilization.

3.2. Inefficiency/Inflexibility of Logical Interface Bandwidth Allocation

Logically-separate routing instances in some implementations further complicates the situation. Dedicating separate physical backbone links, or in the case of sharing of a single common link, dedicating a portion of the link, to each routing instance is not efficient. For example, if there are 2 routing instances and 3 parallel links and half of each link bandwidth is assigned to a routing instance,

then neither routing instance can support an LSP with bandwidth

greater than half the link bandwidth. The same problem is also present in the case of sharing of a single common link using the dedicated logical interface and link bandwidth method. An alternative in dealing with multiple parallel links is to assign a logical interface and bandwidth on each of the parallel physical links to each routing instance, which improves efficiency as compared to dedicating physical links to each routing instance.

Note that the traffic flows and LSPs from these different routing instances effectively operate in a Ships-in-the-Night mode, where they are unaware of each other. Inflexibility results if there are multiple sets of LSPs (e.g., from different routing instances) sharing one link or a set of parallel links, and at least one set of LSPs can preempt others, then more efficient sharing of the link set between the routing instances is highly desirable.

4. Composite Transport Group Framework

4.1. CTG Framework

Composite Transport Group (CTG) is the TE method to transport aggregated traffic over a composite link. A composite link defined in ITU-T [ITU-T G.800] is a single link that bundles multiple parallel links between the two same subnetworks. Each component link in a composite link is independent in the sense that each component link is supported by a separate server layer trail that can be implemented by different transport technologies such as wavelength, Ethernet PHY, MPLS(-TP). The composite link conveys communication information using different server layer trails thus the sequence of symbols across this link may not be preserved.

Composite Transport Group (CTG) is primarily a local traffic engineering and transport framework over multiple parallel links or multiple paths. The objective is for a composite link to appear as a virtual interface to the connected routers. The router provisions incoming traffic over the virtual interface. CTG creates CTG connection and map incoming traffic CTG connections. CTG connections are transported over parallel links, i.e. component links in a composite link. The CTG distribution function can locally determine which component link CTG connections should traverse over. The CTG framework is illustrated in Figure 1 below.

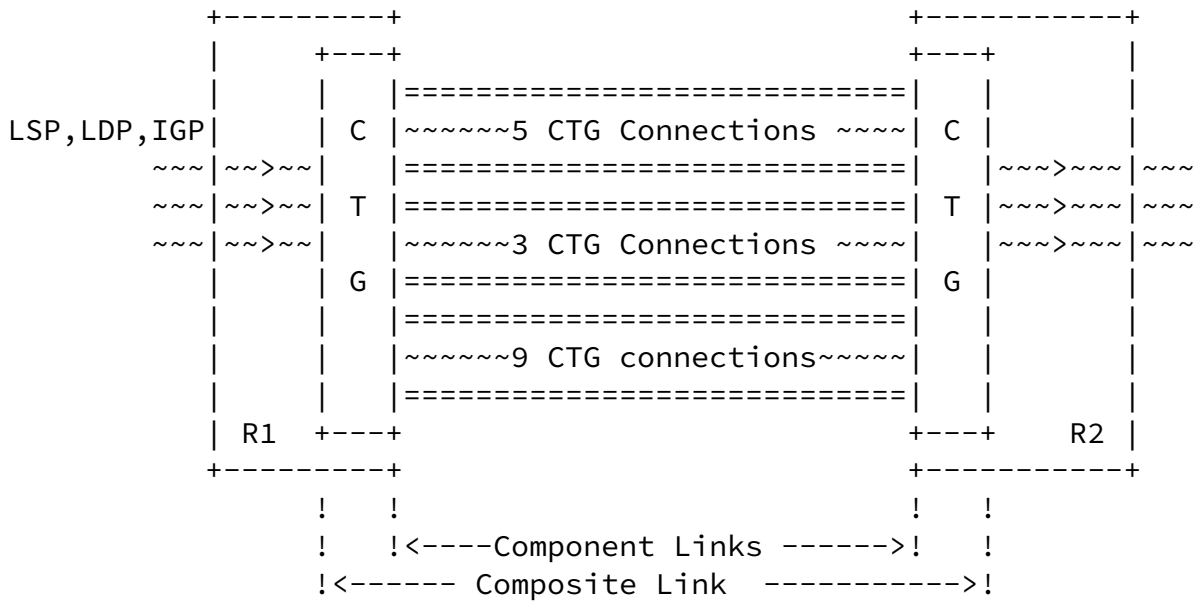

Figure 1: Composite Transport Group Architecture Model

In Figure 1, a composite link is configured between router R1 and R2. The composite link has three component links. To transport LSP traffic, CTG creates a CTG connection for the LSP first, and select a component link to carry the connection. (apply for LDP and IGP traffic as well). A CTG connection only exists in the scope of a composite link. The traffic in a CTG connection is transported over a single component link.

The model in Figure 1 applies two basic scenarios but is not limited to. First, a set of physical links connect adjacent (P) routers. Second, a set of logical links connect adjacent (P or PE) routers over other equipment that may implement RSVP-TE signaled MPLS tunnels, or MPLS-TP tunnels.

A CTG connection is a point-to-point logical connection over a composite link. The connection rides on component link in a one-to-one or many-to-one relationship. LSPs map to CTG connections in a one-to-one or many-to-one relationship. The connection can have the following traffic engineering parameters:

- bandwidth over-subscription
- factor placement
- priority
- holding priority

CTG connection TE parameters can be mapped directly from the LSP parameters signaled in RSVP-TE or can be set at the CTG management interface (CTG Logical Port). The connection bandwidth shall be set. If a LSP has no bandwidth information, the bandwidth will be calculated at CTG ingress using automatic bandwidth measurement function.

LDP LSPs can be mapped onto the connections per LDP label. Both outer label (PE-PE label) and Inner label (VRF Label) can be used for

the connection mapping. CTG connection bandwidth shall be set through auto-bandwidth measurement function at the CTG ingress. When the connection bandwidth tends to exceed the component link capacity, CTG is able to reassign the flows in one connection into several connections and assign other component links for the connections without traffic disruption.

A CTG component link can be a physical link or logical link (LSP Tunnel [LSP Hierarchy]) between two routers. When component links

are physical links, there is no restriction to component link type, bandwidth, and performance objectives (e.g., RTD and Jitter). Each component link maintains its own OAM. CTG is able to get component link status from each link and take an action upon component link status changes.

Each component link can have its own Component Link Cost and Component Link Bandwidth as its associated engineered parameters. CTG uses component link parameters in the assignment of CTG connections to component links.

CTG provides local traffic engineering management over parallel links based on CTG connection TE information and component link parameters. Component link selection for CTG connections is determined locally and may change without reconfiguring the traffic flows. Changing the selection may be triggered by a component link condition change, configuration of a new traffic flow or modification on existing one, or operator required optimization process. The assignment of CTG connections to component links enables TE based traffic distribution and link failure recovery with much less link capacity than current methods mentioned in the section of the problem statements.

CTG connections are created for traffic management purpose on a composite link. They do not change the forwarding schema. The forwarding engine still forwards based on the LSP label created per traffic LSP. Therefore, there is no change to the forwarding.

CTG techniques applies to the situation that the rate of the distinct traffic flows are not higher than the capacity of any component link in composite link.

4.2. CTG Performance

Packet re-ordering when moving a CTG connection from one component link to another can occur when the new path is shorter than the previous path and the interval between packet transmissions is less than the difference in latency between the previous and the new paths. If the new path is longer than the previous path, then re-ordering will not occur, but the inter-packet delay variation will be increased for those packets before and after the change from the previous to the new path. Requirements are stated in this draft to allow an operator to control the frequency of CTG path changes to control the rate of occurrence for these reordering or inter-packet delay variation events.

In order to prevent packet loss, CTG must employ make-before-break when a connection to component link mapping change has to occur. When CTG determines that the current component link for the

connection is no longer sufficient based on the connection bandwidth requirement, CTG ingress establishes a new connection with increased bandwidth on the alternative component link, and switches the traffic onto the new connection before the old connection is torn down. If the new connection is placed on a link that has equal or longer latency than the previous link, the packet re-ordering problem does not occur, but inter-packet delay variation will increase for a pair of packets. When a component link fails, CTG may also move some impacted CTG connections to other component links. In this case, a short service disruption may occur, similar to that caused by other local protection methods.

Time sensitive traffic can be supported by CTG. For example, when some traffic which is very sensitive to latency (as indicated by pre-set priority bits (i.e., DSCP or Ethernet user priority) is being carried over CTG that consists of component links that cannot support the traffic latency requirement, the traffic flow with strict latency requirement can be mapped onto certain component links manually or by using pre-defined policy setting at CTG ingress.

4.3. Differences between CTG and A Link Bundle

4.3.1. Virtual Routable Link vs. TE Link

CTG is a data plane transport function over a composite link. A

composite link contains multiple component links that can carry traffic independently. CTG is the method to transport aggregated traffic over a composite link. The composite link appears as a single routable virtual interface between the connected routers. The component links in composite link do not belong to IGP links in OSPF/IS-IS. The network only maps LSP or LDP to a composite link, i.e. not to individual component links. CTG ingress will select component link for individual LSP and LDP and merge them at composite link egress. CTG ingress does not need to inform CTG egress which component link CTG connections traverse over.

A link bundle [RFC4201] is a collection of TE links. It is a logical construct that represents a way to group/map the information about certain physical resources that interconnect routers. The purpose of link bundle is to improve routing scalability by reducing the amount of information that has to be handled by OSPF/IS-IS. Each physical links in the link bundle are an IGP link in OSPF/IS-IS. A link bundle only has the significance to router control plane. The mapping of LSP to component link in a bundle is determined at LSP setup time and this mapping does not change due to new configurations of LSP/LDP traffic. A link bundle only applies to RSVP-TE signaled traffic, CTG applies to RSVP/RSVP-TE/LDP signaled traffic.

4.3.2. Component Link Parameter Independence

CTG allows component links to have different costs, traffic engineering metric and resource classes. CTG can derive the virtual interface cost from component link costs based on operator policy. CTG can derive the traffic engineering parameter for a virtual interface from its component link traffic engineering parameters.

A Link Bundle requires that all component links in a bundle to have the same traffic engineering metric, and the same set of resource classes.

5. Composite Transport Group Requirements

Composite Transport Group (CTG) is about the method to transport aggregated traffic over multiple parallel links. CTG can address the problems existing in today IP/MPLS network. Here are some CTG requirements:

5.1. Composite Link Appearance as a Routable Virtual Interface

The carrier needs a solution where multiple routing instances see a separate "virtual interface" to a shared composite link composed of

parallel physical/logical links between a pair of routers.

CTG would communicate parameters (e.g., admin cost, available bandwidth, maximum bandwidth, allowable bandwidth) for the "virtual interface" associated with each routing instance.

The "virtual interface" shall appear as a fully-featured routing adjacency in each routing instance, not just an FA [RFC3477] . In particular, it needs to work with at least the following IP/MPLS control protocols: OSPF/IS-IS, LDP, IGP-TE, and RSVP-TE.

CTG SHALL accept a new component link or remove an existing component link by operator provisioning or in response to signaling at a lower layer (e.g., using GMPLS).

CTG SHALL be able to derive the admin cost and TE metric of the "virtual interface" from the admin cost and TE metric of individual component links.

A component link in CTG SHALL be supportable numbered link or unnumbered link in the IGP.

5.2. CTG mapping of Traffic Flows to Component Links

The objective of CTG is to solve the traffic sharing problem at a virtual interface level by mapping LSP traffic to component links (not using hashing):

1. using TE information from the control planes of the routing instances attached to the virtual interface when available, or

2. using traffic measurements when it is not.

CTG SHALL map traffic flows to CTG connections and place an entire connection onto a single component link.

CTG SHALL support operator assignment of traffic flow to component

link.

5.2.1. Mapping Using Router TE information

CTG SHALL use RSVP-TE for bandwidth signaled by a routing instance to explicitly assign TE information to the CTG connection that the LSP is mapped to.

CTG SHALL be able to receive, interpret and act upon at least the following router signaled parameters: minimum bandwidth, maximum bandwidth, preemption priority, and holding priority and apply them to the CTG connections where the LSP is mapped.

5.2.2. Mapping When No Router TE Information is Available

CTG SHALL map LDP-assigned labeled packets based upon local configuration (e.g., label stack depth) to define a CTG connection that is mapped to one of the component links in the CTG.

CTG SHALL map LDP-assigned labeled packets that identify the source-destination LER as a CTG connection.

CTG SHOULD support entropy labels [Entropy Label] to map more granular flows to CTG connections.

In a mapping case, the CTG SHALL be able to measure the bandwidth actually used by a particular connection and derive proper TE information for the connection.

CTG SHALL support parameters that define at least a minimum bandwidth, maximum bandwidth, preemption priority, and holding priority for connections without TE information.

5.3. Bandwidth Control for Connections with and without TE information

The following requirements apply to a virtual interface with CTG capability that supports the traffic flows with TE information and the flows without TE information.

A "bandwidth shortage" issue can arise in CTG if the total bandwidth of the connections with provisioned TE information and those with auto measured TE information exceeds the bandwidth of the composite link.

CTG SHALL support a policy based preemption capability such that, in the event of such a "bandwidth shortage", the signaled or configured preemption and holding parameters can be applied to the following treatments to the connections:

- For a connection that has RSVP-TE LSP(s), signal the router that the LSP has been preempted. CTG SHALL support soft preemption (i.e., notify the preempted LSP source prior to preemption). [Soft Preemption]

- For a connection that has LDP(s), where the CTG is aware of the LDP signaling involved to the preempted label stack depth, signal release of the label to the router

- For a connection that has non-re-routable RSVP-TE LSP(s) or non-releasable LDP(s), signal the router or operator that the LSP or LDP has been lost.

5.4. CTG Transport Resilience

Component links in CTG may fail independently. The failure of a component link may impact some CTG connections. The impacted CTG connections SHALL be replaced to other active component links by using the same rules as of the assignment of CTG connection to component link.

CTG component link recovery scheme SHALL perform equal to or better than existing local recovery methods. A short service disruption may occur during the recovery period.

5.5. CTG Operational and Performance

CTG requires methods to dampen the frequency of connection bandwidth change and/or connection to component link mapping changes (e.g., for re-optimization). Operator imposed control policy SHALL be allowed.

CTG SHALL support latency sensitive traffic.

The determination of latency sensitive traffic SHALL be determined by any of the following methods:

- Use of a pre-defined local policy setting at CTG ingress

- A manually configured setting at CTG ingress

- MPLS traffic class in a RSVP-TE signaling message

The determination of latency sensitive traffic SHOULD be determined (if possible) by any of the following methods:

- Pre-set bits in the Payload (e.g., DSCP bits for IP or Ethernet user priority for Ethernet payload)

6. Security Considerations

CTG is a local function on the router to support traffic engineering management over multiple parallel links. It does not introduce a security risk for control plane and data plane.

7. IANA Considerations

IANA actions to provide solutions are for further study.

8. Acknowledgements

Authors would like to thank Adrian Farrel from Olddog, Ron Bonica from Juniper, Nabil Bitar from Verizon, and Eric Gray from Ericsson for the review and great suggestions.

9. References

9.1. Normative References

[ITU-T G.800]
ITU-T Q12, "Unified Functional Architecture of Transport Network", ITU-T G.800, February 2008.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, March 1997.

[RFC3477] Kompella, K., "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.

[RFC4090] Pan, P., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

[RFC4201] Kompella, K., "Link Bundle in MPLS Traffic Engineering",

RFC 4201, March 2005.

9.2. Informative References

[Entropy Label]
Kompella, K. and S. Amante, "The Use of Entropy Labels in MPLS Forwarding", November 2008, <http://www.ietf.org/internet-drafts/draft-kompella-mpls-entropy-label-01>.

[LSP Hierarchy]
Shiomoto, K. and A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", November 2008, <http://www.ietf.org/internet-drafts/draft-ietf-ccamp-lsp-hierarchy-bis-05.txt>.

[Soft Preemption]
Meyer, M. and J. Vasseur, "MPLS Traffic Engineering Soft Preemption", February 2009, <http://www.ietf.org/internet-drafts/draft-ietf-mpls-soft-preemption-16.txt>.

Authors' Addresses

So Ning
Verizon
2400 N. Glem Ave.,
Richerson, TX 75082

Phone: +1 972-729-7905
Email: ning.so@verizonbusiness.com

Andrew Malis

Verizon
117 West St.
Waltham, MA 02451

Phone: +1 781-466-2362
Email: andrew.g.malis@verizon.com

Dave McDysan
Verizon
22001 Loudoun County PKWY
Ashburn, VA 20147

Phone: +1 707-886-1891
Email: dave.mcdysan@verizon.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075

Phone: +1 469-229-5387
Email: lucyyong@huawei.com

Frederic Jounay
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex,
FRANCE

Phone:
Email: frederic.jounay@orange-ftgroup.com