Francois Le Faucheur,
Cisco Systems, Inc.

IETF Internet Draft
Expires: December, 2002
Document: draft-lefaucheur-tewg-russian-dolls-00.txt June, 2002

Considerations on Bandwidth Constraints Models for DS-TE

Status of this Memo

Abstract

This document provides input for the selection of a default Bandwidth Constraints Model for Diff-Serv-aware MPLS Traffic Engineering (DS-TE).

It discusses a number of considerations on Bandwidth Constraints Models and how the Maximum Allocation Model and the Russian Dolls Model address these considerations.

While this document may not exhaustively cover all possible considerations for selection of a Bandwidth Constraints model, we feel it covers the most important considerations for practical DS-TE deployment.

We conclude that the Russian Dolls Bandwidth Constraint Model is a

good default Bandwidth Constraint Model for DS-TE.

1. Introduction

[DSTE-REQ] presents the Service Providers requirements for support of Diff-Serv-aware MPLS Traffic Engineering (DS-TE). [DSTE-REQ] states that a default Bandwidth Constraints Model must be specified as part of the DS-TE solution. The purpose of such a default model is to ensure that there is at least one common Bandwidth Constraints model implementation across various vendors equipment in order to allow for easier deployment of DS-TE.

Note that additional Bandwidth Constraints models may also be specified and supported by DS-TE implementations.

2. Terminology

Section 3.3 of [DSTE-REQ] describes two examples of Bandwidth Constraints Models.

The first example uses a separate, independent Bandwidth Constraint (BC) for each Class-type (CT). We refer to this model as the Maximum Allocation Model or MAM.

With MAM, the Bandwidth Constraints are defined in the following manner:

All LSPs supporting Traffic Trunks from CTc use no more than BCc

For example, when 3 CTs are used with MAM:

- sum (CT0) <= BC0
- sum (CT1) <= BC1
- sum (CT2) <= BC2

For illustration purposes, on a link of 100 unit of bandwidth where three CTs are used, the network administrator might then configure

BC0=30, BC1= 50, BC2=20 such that:

- All LSPs supporting Traffic Trunks from CT0 use no more than 30 (e.g. Voice <= 30)
- All LSPs supporting Traffic Trunks from CT1 use no more than 50 (e.g. Premium Data <= 50)
- All LSPs supporting Traffic Trunks from CT2 use no more than 20 (e.g. Best Effort <= 20)

The second example is the Russian Dolls Model. We refer to it as the RDM. More details can also be found on the RDM in section 9 and Appendix C of [DSTE-PROTO].

With RDM, the Bandwidth Constraints are defined in the following manner:

BCb is the constraint that bounds the total bandwidth used by all LSPs supporting Traffic Trunks from all class types CTn, where b <= n < M, M being the number of class-types used in the network.

For example, when 3 CTs are used with RDM (M=3):

- sum (CT0+CT1+CT2) <= BC0
- sum (CT1+CT2) <= BC1
- sum (CT2) <= BC2

For illustration purposes, on a link of 100 units of bandwidth where three CTs are used, the network administrator might then configure BC0=100, BC1= 80, BC2=60 such that

- All LSPs supporting Traffic Trunks from CT2 use no more than 60 (e.g. Voice <= 60)
- All LSPs supporting Traffic Trunks from CT1 or CT2 use no more than 80 (e.g. Voice + Premium Data <= 80)
- All LSPs supporting Traffic Trunks from CT0 or CT1 or CT2 use no more than 100 (e.g. Voice + Premium Data + Best Effort <= 100).

3. Considerations

3.1. Canonical DS-TE Deployment

For easier discussion of the considerations below, we consider an example DS-TE deployment which we refer to as the canonical DS-TE deployment.

The canonical DS-TE deployment is characterized by:

- 3 Class-Types :
  - o CT2 for real-time PHB Scheduling Class(es) (PSCs; see [DIFF-NEW])
  - o CT1 for low-loss PSCs
  - o CT0 for other PSCs (e.g. Best Effort)
- CT2 needs high preemption priority(ies) to ensure placement of such traffic as close as possible to its shortest path.
- CT1 needs medium preemption priority(ies) to ensure CT1 traffic is as close as possible to its shortest path, but without forcing some CT2 traffic further away from its own shortest path.
- CT0 only needs low preemption priority(ies) as its QoS objectives can be accommodated, if necessary, by paths relatively further away from their shortest path as long as they satisfy the bandwidth/resources constraints.

Note that although we always refer to the canonical DS-TE deployment in the discussion below, the discussion also applies to other deployment scenarios.

3.2. Canonical Set of DS-TE Objectives

The considerations below stem from a set of typical practical objectives sought in DS-TE deployment scenarios. We refer to those as the canonical set of DS-TE objectives. This set includes:

- Diff-Serv QoS enforcement. We wish to use the DS-TE Bandwidth Constraints to ensure the respective QoS performance targets of the various Diff-Serv Behavior Aggregates are always met

regardless of the actual demand of LSPs across all CTs.

- Avoiding bandwidth wastage. If bandwidth is not used to establish LSPs of a given CT, this bandwidth should be available for use by other CTs, as long as it does not compromise the previous objective. This is also referred to as achieving efficient bandwidth sharing across CTs. Note that we are talking about use of bandwidth in the control plane for Constraint Based Routing and not about use of bandwidth in the data plane. Diff-Serv PHB implementations are responsible for achieving efficient bandwidth sharing in the data plane, e.g. through use of work-conserving scheduling algorithms).

- Avoiding starvation of Best-Effort traffic (considering that when preemption is used, Best-Effort LSPs are typically granted low preemption priority).

3.3. Avoiding Wastage and QoS Degradation simultaneously

MAM can not simultaneously protect against both bandwidth wastage and QoS degradation. With MAM:

- EITHER, one configures Bandwidth Constraints so that that SUM (BCi) <= link bandwidth.

Then, significant bandwidth wastage can occur because whenever a CT is not using its bandwidth, this bandwidth cannot be used by other CTs.
Consider the example where the link has 100 units of bandwidth and BC0=35, BC1=35 and BC2=30. If, Sum (bandwidth of all LSPs of CT2)= 10, then LSPs of CT0 are still limited to 35 and LSPs of CT1 to 35, so that 20 units of bandwidth will go wasted.

- OR, one configures Bandwidth Constraints so that SUM (BCi) exceeds link bandwidth.

Then, constraint-based routing and admission control can not protect against aggregate congestion and associated QoS degradation.
Consider the example where a link has 100 units of bandwidth and BC0=70, BC1=70 and BC2=50. Then, the router performing admission control for that link will accept establishment of LSPs whose sum across all classes can reach 190, far exceeding the link capacity. Note that this could result in QoS degradation not just on CT0 LSPs but also on CT1 LSPs (for example if there are 60 units of CT1 LSP established and 50 units of CT2 LSPs, which totals to 110% of link capacity, it is clear that CT1 will experience QoS degradation. (Depending on the scheduler used, CT2 might also experience degradation.)

RDM, by contrast, can simultaneously protect well against bandwidth wastage (i.e. achieve efficient bandwidth sharing across CTs) and protect against QoS degradation. This is because the recursiveness of Bandwidth Constraints always allows a CT to make use of what is left over by the previous CT. For example, whatever is unused by CT2 can be reused by CT1, whatever is unused by CT1 and CT2 will be reused by CT0 since CT0 is only constrained collectively with CT1 and CT2 by BC0. Yet RDM can avoid QoS degradation by separately constraining the amount of real-time traffic as well as the amount of real-time plus low-loss traffic (see also section 3.5 below). We note that RDM does not completely remove the possibility of conceivable bandwidth wastage. However, we also observe that:

- it considerably reduces this risk, as compared to MAM, by effectively constraining CTs collectively and thus always giving the opportunity to reuse the unused bandwidth to all CTs with smaller numerical indexes. So even if it does not give such opportunity to all other CTs (which would be the ideal if bandwidth wastage were the only concern) it does provide many opportunities for bandwidth reuse.
- the remaining conceivable bandwidth wastage scenarios in RDM may be more or less inevitable anyway to meet QoS objectives of Diff-Serv classes. Therefore these scenarios do not really represent bandwidth wastage due to the BC model itself. For example, a conceivable bandwidth wastage scenario with RDM is when there is little CT0 demand, and a heavy CT1 and CT2 demand. Bandwidth may be considered wasted since some CT1/CT2 demand will get rejected even if link is not fully used (since CT2+CT1 is limited by BC1 below link capacity). But limiting CT1 and CT2 to less than link capacity collectively, even when there is no CT0 traffic, is probably necessary anyway to ensure the QoS objectives of low-delay and low-loss traffic are met. So such a bandwidth wastage can be seen as largely inevitable since accepting more CT1/CT2 traffic would eventually result in QoS degradation. In fact, such "bandwidth wastage" may be seen as desirable in order to avoid QoS degradation.

3.4. Avoiding Starvation of Best-Effort Traffic

In typical DS-TE deployment scenario, MAM does not effectively

protect low priority traffic (ie CT0) against starvation.

With MAM, in order to avoid the bandwidth wastage issue pointed out above, BC1 and BC2 need to be configured as high as possible. Yet to avoid QoS degradation of CT1 and CT2, BC1 and BC2 need to be configured so that BC1+BC2 is below link capacity. So, it is expected that in practise, BC1 and BC2 would be commonly configured so that BC1+BC2 is just slightly below link capacity - say to "capacity minus small-delta". Since, when preemption is used, CT0 traffic is typically granted low preemption priorities (to maximize QoS performance of CT1 and CT2 traffic), whenever CT1 and CT2 traffic are grabbing all their allowed resources, CT0 traffic will be starved out

and left with only "small-delta" units of bandwidth. Increasing the value of "small-delta" to alleviate CT0 starvation could be done, but this would be at the cost of increasing bandwidth wastage.

With RDM, low priority traffic (i.e. CT0) may be well protected against starvation, regardless of the preemption priority it uses, by setting BC1 (which constrains CT1 + CT2 traffic) to less than link capacity, thus ensuring that the required capacity is left for CT0.

3.5. Diff-Serv QoS Enforcement Objective

DS-TE allows enforcement of different Bandwidth Constraints. These multiple Bandwidth Constraints can be useful to pursue multiple different goals.

One such goal is the "Diff-Serv QoS enforcement objective" identified above in the set of canonical DS-TE objectives. This objective is to ensure that the respective QoS performance targets for the Diff-Serv PSC(s) belonging to each CT are met. In other words, the Bandwidth Constraints are used to control the distribution of traffic across the various CTs so that the traffic load submitted to the various PHBs/PSCs activated on the links is compatible with their respective targeted QoS performance. In that context, DS-TE can be seen as a form of aggregate admission control for Diff-Serv.

Other goals relate more to how resources can be apportioned across different classes in order to address some Service Provider policy. We refer to such goals as "Policy goals". An example of a policy goal would be the desire to limit the amount of traffic that Best Effort traffic may be able to use on a given link to a certain level (even if going beyond that level would not result in degradation of the QoS objectives).

We feel that, while addressing policy goals is highly desirable, addressing the Diff-Serv QoS enforcement objective well is of paramount importance, as this is the most fundamental thrust behind DS-TE (ie being Diff-Serv-aware as opposed to being aware of generic policy classes).

We feel that RDM is a much more natural match to the Diff-Serv QoS enforcement objective than MAM because:

- when there is no CT1 and CT2 traffic, there is typically no need to limit CT0 traffic below "link capacity" in order to meet CT0 traffic QoS objectives.
  - RDM will not unnecessary limit CT0 traffic when there is no CT1 and CT2 traffic since the only constraint applying to CT0 is that CT0+CT1+CT2 <= BC0.
  - To avoid unnecessarily limiting CT0 when there is no CT1 and CT2 traffic, MAM would have to have its BC0 configured to "link capacity". This means that BC0+BC1+BC2 would significantly exceed link capacity

resulting in significant QoS degradation whenever there is CT1 and/or CT2 traffic.

- When there is some CT1 and CT2 traffic, it is typically useful to effectively limit CT0 to whatever is left over by CT1 and CT2 from the link capacity, in order to maintain CT0 traffic QoS objectives (since the CT1/CT2 PHBs will typically be configured so that they are granted the bandwidth/resources they require for CT1 and CT2 traffic).
  - RDM naturally limits CT0 to whatever is left by CT1 and CT2 since CT0 is effectively limited by BC0-CT1-CT2.

    - To limit CT0 to whatever is left over by CT1 and CT2 in all cases, MAM would have to configure BC0 to a value smaller than ("link capacity"-BC1-BC2) which would be very small if not equal to zero, unless significant bandwidth wastage is tolerated.
- Similarly, intuition (as well as some experimental observations) suggests that it is efficient to collectively limit CT1 and CT2. This enables CT1 to effectively make full use of whatever bandwidth hasn't been used by CT2 to avoid bandwidth wastage, while protecting CT1 traffic from QoS degradation by constraining CT1 more tightly as the amount of CT2 traffic increases. In simple terms, if there is less real-time (CT2) traffic established on the link, the link can take up more low-loss (CT1) traffic without QoS degradation of the low loss traffic (assuming appropriate configuration of the PHBs on the link or assuming dynamic adjustment of the PHBs).
    - RDM naturally limits CT1 and CT2 collectively via BC1.
    - MAM cannot naturally limit CT1 depending on the amount of CT2 traffic.

4. Conclusions

Considering that:

- other Bandwidth Constraints Models can be defined in addition to the default model to address less typical/more complex deployment scenarios
- the Russian Dolls Model matches very well the canonical DS-TE objectives in the canonical DS-TE deployment scenario (as well as many other practical deployment scenarios)

we recommend selecting the Russian Dolls Model as the default model for DS-TE.

5. Security Considerations

No new security considerations are raised by this document. Those are the same as the ones mentioned in [DSTE-REQ].

6. Acknowledgments

We thank Bruce Davie for his review and recommendations.

References

[DSTE-REQ] Le Faucheur et al, Requirements for support of Diff-Serv-aware MPLS Traffic Engineering, draft-ietf-tewg-diff-te-reqts-05.txt, June 2002.

[DSTE-PROTO] Le Faucheur et al, Protocol extensions for support of Diff-Serv-aware MPLS Traffic Engineering, draft-ietf-tewg-diff-te-proto-01.txt, June 2002.

[DIFF-NEW] Grossman, " New Terminology and Clarifications for Diffserv ", RFC3260, April 2002.

Author's Address:

Francois Le Faucheur
Cisco Systems, Inc.
Village d'Entreprise Green Side - Batiment T3
400, Avenue de Roumanille
06410 Biot-Sophia Antipolis
France
Phone: +33 4 97 23 26 19
Email: flefauch@cisco.com