Network Working Group
Internet Draft
<draft-ash-mpls-diffserv-te-alternative-02.txt>
Category: Informational
Expiration Date: February 2002

Jerry Ash
Luyuan Fang
Wai Sum Lai
AT&T

August 2001

Alternative Technical Solution for MPLS DiffServ TE

Status of this Memo

Abstract

Service Provider requirements for support of DiffServ-aware MPLS Traffic Engineering (DS-TE) are presented in [DS-TE-REQ]. [DS-TE-SOLN] describes a proposed technical solution for meeting the DS-TE requirements. The draft proposes complex IGP extensions of per-class-type link-state advertisements (LSA) to communicate per-class-type available bandwidth, etc. There is concern about scalability of the IGP overhead, and particularly the IGP response to overloads and failures [IGP-SCALE]. This Draft presents an alternative technical solution which avoids further extensions of the IGP. We give an example of this in this draft which shows how to use measurement-based/reservation-crankback admission control rather than flooding more per-CT-available-bandwidth information). This draft proposes this as an alternative Technical Solution for discussion.

Table of Contents

1. Introduction

Service Provider requirements for support of DiffServ-aware MPLS Traffic Engineering (DS-TE) are presented in [DS-TE-REQ]. DS-TE is discussed in the Traffic Engineering Working Group Framework document [TEWG-FW]. DS-TE requirements are defined for class types (CTs), where CTs are defined in [TEWG-FW] as aggregations of individual service classes. Instead of having per-class parameters being configured and propagated on each LSR interface, classes are aggregated into CTs having common per-CT parameters (e.g., minimum bandwidth) to satisfy required performance levels, however, no bandwidth requirements are enforced for classes within a CT. The main motivation for grouping a set of classes into a CT is to improve the scalability of IGP LSAs by propagating information on a per-CT basis instead of on a per-class basis, and also to allow better bandwidth sharing between classes in the same CT.

[DS-TE-SOLN] describes a proposed technical solution for meeting the DS-TE requirements. The draft proposes complex IGP extensions of per-class-type link-state advertisements (LSA) to communicate per-class-type available bandwidth, etc. It already gets so complex that the draft proposes to compress the advertisements. There is concern about scalability of the IGP overhead, and particularly the IGP response to overloads and failures [IGP-SCALE]. Hence there is concern about further significant extensions to increase IGP overhead, which will further exacerbate the problem identified in [IGP-SCALE]. Furthermore we think the extensions are unnecessary, because there are other, equally effective ways to do DS-TE without the IGP TE extensions.

The draft addresses both establishing an LSP and modifying LSPs (e.g., increasing LSP bandwidth allocation, such as described in [MODIFY]. Either of these functions might use additional IGP/LSA advertisements/fields in order to update the TE database (TED) so as to select an LSP path to establish and/or modify. The draft points out that such additional IGP/LSA extensions are not necessarily required if one uses alternative methods to select/modify LSPs on a per-CT basis, such as through use of LSP event-dependent-routing/crankback methods [QoS-ROUTING], which is illustrated by an example. The example illustrates that the per-CT extension for IGP/LSA's described in [DS-TE-SOLN], is not really necessary. Making the proposed extensions, which only exacerbates the problem and concern with IGP/LSA overhead [IGP-SCALE], can be

avoided.

First we briefly review concerns over the scalability of IGPs, and then present an alternative technical solution which avoids further extensions of the IGP. We give an example of this which shows how to use measurement-based/reservation-crankback admission control rather than flooding more per-CT-available-bandwidth information). This draft proposes this as an alternative Technical Solution for discussion.

2. Concerns over the Scalability of IGP Link-State Protocols

Congestion can arise in data networks for many different reasons. There is evidence based on previous failures that link state (LS) routing protocols, such as OSPF and ISIS, currently can not recover from large failures which result in widespread loss of topology database information (especially when areas/peer-groups get "too large"). LS protocols typically use topology-state update (TSU) mechanisms to build the topology database at each node, typically conveying the topology status through flooding of TSU messages containing link, node, and reachable-address information between nodes. In OSPF, they use the link state advertisement (LSA), in PNNI, such mechanisms use the PNNI topology state element (PTSE), in frame-relay and proprietary-routing networks, they may use other TSU mechanisms to exchange topology status information to build the topology database at each node.

Earlier papers and contributions identified issues of congestion control and failure recovery for LS protocol networks, such as OSPF, ISIS, and PNNI networks [IGP-SCALE, maunder, choudhury, pappalardo1, pappalardo2, atm01-0101]. In [IGP-SCALE] much evidence is presented of the current problems associated with LS failure recovery from various failure conditions, which is based on a) failure experience, b) vendor analysis of product performance, and c) analytic modeling, simulation analysis, and emulation analysis.

As to failure experience, AT&T has experienced serious data network outages in which recovery of the underlying LS protocols was inadequate. For example, in the failure in the AT&T Frame Relay Network on April 13, 1998 [att], an initial procedural error triggered two undetected software bugs, leading to a huge overload of control messages in the network. The result of this control overload was the loss of all topology database information, and the link-state protocol then attempted to recover the database with the usual Hello and TSU updates.

Analysis has shown that several problems then occurred to prevent the network from recovering properly:

- Very large number of TSUs being sent to every node to process, causing general processor overload

- Route computation based on incomplete topology recovery, causing routes to be generated based on transient, asynchronous topology information and then in need of frequent re-computation

- Inadequate work queue management to allow processes to complete before more work is put into the process queue

- Inability to segment the network (into smaller "peer groups") to aid in the link-state protocol recovery

- Inability to access the node processors with network management commands due to lack of necessary priority of these messages

A more recent failure occurred on February 20, 2001 in the AT&T ATM Network, which resulted in a large overload of TSUs, and a lengthy network outage [pappalardo1, pappalardo2]. Manual procedures were put in place to reduce TSU flooding, which worked to stabilize the network. It is desirable that such TSU flooding reduction be automatic under overload.

In general, there have been a number of major outages reported by most major carriers, and routing protocol issues have generally been involved. Other relevant LS-network failures are reported in [cholewka, jander].

Various networks employing LS protocols use various control messages and mechanisms to update the LS database, not necessarily LSAs, PTSEs, or flooding mechanisms. Based on experience, however, the LS protocols are found to be vulnerable to loss of database information, control overload to re-sync databases, and other failure/overload scenarios which make such networks more vulnerable in the absence of adequate protection mechanisms. Hence we are addressing a generic problem of LS protocols across a variety of implementations, and the basic problem is prevalent in LS protocol networks employing frame-relay, ATM, and IP based technologies.

As a result of these failures, a number of congestion control/failure recovery mechanisms are being recommended [IGP-SCALE]. The goal is to enable LS protocols to a) gracefully recover from massive loss of topology database information, and b) respond gracefully to network overloads and failures. [IGP-SCALE] proposes specific additional considerations for network congestion control/failure recovery. Candidate mechanisms are proposed for control of network congestion and failure recovery, in particular the following mechanisms are proposed for investigation in OSPF and ISIS working groups:

a) throttle new connection setups, topology-state updates, and Hello updates based on automatic congestion control mechanisms,
b) special marking of critical control messages (e.g., Hello and

topology-state-update Ack) so that they may receive prioritized processing,

c) database backup, in which a topology database could be automatically recovered from loss based on local backup mechanisms, and

d) hitless restart, which allows routes to continue to be used if there is an uninterrupted data path, even if the control path is interrupted due to a failure.

There is much work already underway in standards bodies, namely the IETF, ATM Forum, and ITU-T, to address issues of congestion control and failure recovery in ATM- and IP-based packet networks. Numerous references are cited and are further explained in the document [maunder, moy1, moy2, moy3, murphy, whitehead, zinin, atm01-0101, btd-cs-congestion-02.00].

3. DS-TE Technical Solution Alternative

3.1 General Requirements

The following are some proposed, high-level requirements for MPLS-DiffServ TE, which address some of the IGP scalability concerns discussed in Section 2. This is all very preliminary and high-level, and intended to initiate further discussion. Also, the numerical values below are for illustrative purposes only.

1. No new LSAs used to signal per-CT available bandwidth, maximum bandwidth, preemption parameters, etc. Rather, CT-bandwidth allocated and protected by mechanisms that do not require new per-CT LSAs, such as in #3-5 below. (Note that [boyle] proposed that current specifications could be adapted to accomplish MPLS-DiffServ TE.)

2. MPLS connection-admission control and QoS/DiffServ signaling should be decoupled.

3. CT-Bandwidth allocated and protected by MPLS connection admission control, but done without additional LSA extensions.

3.2 Example DS-TE Technical Solution Alternative

We now give an existence proof example of how these requirements can be met. The example presented uses measurement-based/reservation-crankback admission control rather than flooding more per-CT-available-bandwidth information. This draft proposes this as an alternative Technical Solution for discussion.

We now present the details of the example using the following 6 example class-types (CT):

CT 1: QoS class: Y.1541 Class-0, interactive (real-time),
DiffServ EF

```
      admission-control priority: key
      restoration priority: premium
CT 2: QoS class: Y.1541 Class-1, interactive (real-time),
                 DiffServ EF
      admission-control priority: normal
      restoration priority: basic
CT 3: QoS class: Y.1541 Class-3, non-interactive (low loss),
                 DiffServ AF1
      admission-control priority: key
      restoration priority: premium
CT 4: QoS class: Y.1541 Class-2 or 4, non-interactive (low loss),
                 DiffServ AF2
      admission-control priority: normal
      restoration priority: basic
CT 5: QoS class: Y.1541 Class-5, Unspecified, DiffServ BE
      admission-control priority: best-effort
      restoration priority: unprotected
CT 6: QoS class: Y.1541 Class-0, interactive (real-time),
                 control traffic, DiffServ EF
      admission-control priority: key (LSP preemption allowed)
      restoration priority: premium (LSP preemption allowed)
```

In the above CT definitions we generalize the notion of CTs and consider them to be a combination of a) QoS classes (e.g., as specified in [Y.1541] consistent with DiffServ queuing priority classes), b) admission-control priority classes, and c) restoration priority classes at both the MPLS-LSP and transport link level. This is discussed further in a forthcoming Internet Draft. Restoration priority is a way of giving preference to protect higher priority LSPs ahead of lower priority LSPs. A premium service LSP can be protected in preference over a basic service LSP. Admission control priority is a way of giving preference to admit higher priority LSPs ahead of lower priority LSPs. A key service LSP can be admitted in preference over a normal service LSP. For both restoration and admission control, no preemption of existing LSPs is assumed beyond what is specified for CT6.

We now present the details of the example, which is proposed in this draft as an alternative Technical Solution for discussion.

1. CT-Bandwidth allocated and protected by MPLS connection admission control, but done without additional LSA extensions.
   a. at ingress LER:
      - CTs 1 and 3 given unrestricted access to bandwidth on any candidate LSP up to 10% of total traffic load; beyond 10% of total traffic load, bandwidth allocated only when > 5% bandwidth is idle (reservation signaled in the latter cases, perhaps using Setup Priority parameter);
      - CTs 2 and 4 given unrestricted access to bandwidth only on primary LSP up to the protected-CT-bandwidth level; otherwise (on alternate LSPs and/or when protected-CT-

bandwidth exceeded) bandwidth allocated only when > 5% bandwidth is idle (reservation signaled in the latter cases, perhaps using Setup Priority parameter);
- CT 5 allocated up to maximum protected-CT-bandwidth of 1% only on primary LSP, no alternate LSPs allowed;
- ingress LER signals class type (perhaps using L-LSP parameter) and bandwidth allocation to transit LSRs in LSP.

b. at transit LSRs
- bandwidth allocation protected by QoS mechanisms (DiffServ priority, policing, etc.) according to signaled class type;
- reservation not signaled (perhaps using Setup Priority parameter): bandwidth allocation unrestricted, if bandwidth unavailable, crankback to ingress LER;
- reservation signaled (perhaps using Setup Priority parameter): bandwidth allocation restricted to when > 5% bandwidth is idle, if bandwidth unavailable, crankback to ingress LER;

c. CAC is applied for bandwidth allocation per-aggregated-bandwidth-CT, not per microflow.

2. Protected-CT-bandwidth limit can be pre-provisioned per node-pair

3. Protected-CT-bandwidth can be dynamically computed per node-pair, for example:

```
PBWi    = protected bandwidth for CT i
PBWi(w) = .5 x  PBWi(w-1) + .5 x BWIPi(w)
BWIPi   = average bandwidth-in-progress across a load set
          period on CT i
```

The quantities PBWi are computed periodically, such as every week w, per node-pair.

4. MPLS LSP restoration
   a. assigns a minimum of 5 diverse LSP backup path per premium-CT LSP
   b. assigns a minimum of 2 diverse LSP backup path per basic-CT LSP
   c. triggers redirecting all flows to backup LSPs upon specified triggers (e.g., LOS, LOF)
   d. sequentially hunts backup LSPs for available bandwidth to redirect flows
   e. alternatively, dynamically compute and hunt backup LSPs

5. Transport link restoration
   a. assigns a minimum of 5 diverse backup transport paths per premium-CT transport link
   b. assigns a minimum of 2 diverse backup transport paths per premium-CT transport link
   c. triggers redirecting all LSPs to backup transport paths upon specified triggers (e.g., LOS, LOF)
   d. sequentially hunts backup transport paths for available bandwidth to redirect transport links
   e. alternatively, dynamically compute and hunt backup transport

paths

6. No preemption of MPLS-LSPs and/or transport links across CTs, except for control-traffic CT.

The above example addresses both establishing an LSP and modifying LSPs (e.g., increasing LSP bandwidth allocation, such as described in [MODIFY]). That is, the process of adding traffic to an LSP will result in:

1. evaluating whether the LSP (not the topology) has enough capacity.
2. If the LSP does not currently have enough capacity, evaluating whether the topology will permit increasing the capacity.
3. If the topology will not permit increasing the capacity, either re-placing the LSP or establishing a new LSP, using appropriate information to decide where to place it.

Any of these functions might use additional IGP/LSA advertisements/fields in order to update the TED so as to select an LSP path to establish and/or modify. The example points out that such additional IGP/LSA extensions are not necessarily required if one uses alternative methods to select/modify LSPs on a per-CT basis, such as through use of LSP event-dependent-routing/crankback methods [QoS-ROUTING], which is illustrated by an example.

The example illustrates that the per-CT extension for IGP/LSA's described in [DS-TE-SOLN], is not really necessary. Making the proposed extensions, which only exacerbates the problem and concern with IGP/LSA overhead [IGP-SCALE], can be avoided.

3.3 Scalability Comparison

The crankback approach may cause more signaling messages in place of routing information. The scalability comparison between the crankback method and TE routing extensions has been evaluated in some earlier work. There is considerable experience in other networks with such methods [ASH], and there are simulation studies for IP-based networks reported in [QoS-ROUTING] (e.g., see ANNEX 4, Section 4.7). In [QoS-ROUTING], simulation data is presented comparing the scalability between the crankback method and TE routing extensions.

Table 1 gives an example comparison of the performance of state dependent routing (SDR) with LSA flooding compared to event dependent routing (EDR) described in the draft. The numbers in the table give the total messages of each type needed to do the indicated TE functions, including flow setup, bandwidth allocation, crankback, and LSA flooding to update the traffic engineering database (TED). The SDR TE method does available link bandwidth (ALBW) flooding to update the TED while the EDR

method does not. In the simulation there is a 6-times focused overload on one node (OKBR), and clearly the SDR/flooding method is consuming more message resources, particular LSA flooding messages, than the EDR method, while the traffic lost/delayed performance of the two methods is comparable [QoS-ROUTING].

Table 1
Performance Comparison of SDR/flooding Vs. EDR
6X focused overload on OKBK
(total number of messages in simulation)

TE Function	Message Type	SDR/ flooding	EDR
Flow Routing	Flow Setup	18,758,992	18,758,992
QoS Resource Management (LSP Rtg., BW Alloc., Queue Mgmt.)	LSP Bandwidth Allocation	18,469,477	18,839,216
	Crankback	30,459	12,850
TE Database Update	LSA	14,405,040	0

These results, plus experience and the other referenced comparisons favor a method which does not further increase IGP/LSA overhead.

4. Security Considerations

There are no new security considerations based on proposals in this draft.

5. Acknowledgements

The authors gratefully acknowledge the comments and suggestions from many people. At AT&T we thank Chuck Dvorak, Al Morton, and Percy Tarapore, Joel Halpern at Longitude Systems, Lei Yao at Worldcom, and Kwangil Lee at NIST.

6. References

[DS-TE-REQ] Le Faucheur, F., et. al., "Requirements for support of Diff-Serv-aware MPLS Traffic Engineering," work in progress.

[DS-TE-SOLN] Le Faucher, F., et. al., " Protocol extensions for support of Diff-Serv-aware MPLS Traffic Engineering," work in progress.

[BOYLE] Boyle, J., "Accomplishing DiffServ TE Needs with Current Specifications," work in progress.

[KOMPELLA] Kompella, K., "Bandwidth Accounting for Traffic Engineering," work in progress.

[MODIFY] Ash, J., et. al., "LSP Modification Using CR-LDP," work in progress.

[QoS-ROUTING] Ash, J., "Traffic Engineering & QoS Methods for IP-, ATM-, & TDM-Based Multiservice Networks," work in progress.

[TE-REQ] Awduche et al, Requirements for Traffic Engineering over MPLS, RFC2702, September 1999.

[TEWG-FW] Awduche et al, A Framework for Internet Traffic Engineering, work in progress.

[OSPF-TE] Katz, Yeung, Traffic Engineering Extensions to OSPF, work in progress.

[ISIS-TE] Smit, Li, IS-IS extensions for Traffic Engineering, work in progress.

[RSVP-TE] Awduche et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels", work in progress.

[DIFF-MPLS] Le Faucheur et al, "MPLS Support of Diff-Serv", work in progress.

[CR-LDP] Jamoussi et al., "Constraint-Based LSP Setup using LDP", work in progress.

[DIFF-NEW] Grossman, "New Terminology for Diffserv", work in progress, work in progress.

[IGP-SCALE] Ash, G., et. al., Proposed Mechanisms for Congestion Control/Failure Recovery in OSPF & ISIS Networks, work in progress.

[af-pnni-0055.000] "Private Network-Network Interface Specification Version 1.0 (PNNI 1.0)," March 1996.

[ash] Ash, G. R., "Dynamic Routing in Telecommunications Networks," McGraw Hill.

[atmf00-0249] "Scalability and Reliability of large ATM networks."

[atm00-0257] "Signaling Congestion Control in PNNI Networks: The Need and Proposed Solution Outline."

[atm00-0480] "Congestion Control/Failure Recovery in PNNI Networks."

[atm01-0101] "Proposed Mechanisms for Congestion Control/Failure Recovery in PNNI Networks."

[att] "AT&T announces cause of frame-relay network outage," AT&T Press Release, April 22, 1998.

[btd-cs-congestion-02.00] "Signaling Congestion Control Version 1.0", Baseline Text

[cholewka] Cholewka, K., "MCI Outage Has Domino Effect," Inter@ctive Week, August 20, 1999.

[choudhury] Choudhury, G., Maunder, A. S., Sapozhnikova, V., "Faster Link-State Convergence and Improving Network Scalabiity and Stability," sumitted for presentation at LCN 2001.

[hosein1] Hosein, P., "An Improved ACC Algorithm for Telecommunication Networks," Telecommunication Systems 0, 1998.

[hosein2] Hosein, P., "Overload Control for Real-Time Telecommunication Databases," International Teletraffic Congress - 16, Edinburgh, Scotland, June 1999.

[jander] Jander, M., "In Qwest Outage, ATM Takes Some Heat," Light Reading, April 6, 2001.

[maunder] Maunder, A. S., Choudhury, G., "Explicit Marking and Prioritized Treatment of Specific IGP Packets for Faster IGP Convergence and Improved Network Scalability and Stability," draft-ietf-ospf-scalability-00, March 2001.

[mummert] Mummert, V. S., "Network Management and its Implementation on the No. 4ESS," International Switching Symposium, Japan, 1976.

[moy1] Moy, J., "Hitless OSPF Restart", work in progress.

[moy2] Moy, J., "Flooding over parallel point-to-point links," work in progress.

[moy3] Moy, J., "Flooding Over a Subset Topology," work in progress.

[murphy] Murphy, P., "OSPF Floodgates," work in progress.

[pappalardo1] Pappalardo, D., "Can one rogue switch buckle AT&T's network?," Network World Fusion, February 23, 2001.

[pappalardo2] Pappalardo, D., "AT&T, customers grapple with ATM net outage," Network World, February 26, 2001.

[Q.764] "Signalling System No. 7 - ISDN user part signalling procedures," December 1999.

[whitehead] Whitehead, Martin, "A class of overload controls based on controlling call reject rates," ITU-T contribution D.19, Feburary 2001.

[zinin] Zinin, A., et. al., "OSPF Restart Signaling," work in progress.

7. Authors' Addresses

Jerry Ash
AT&T
Room MT D5-2A01
200 Laurel Avenue
Middletown, NJ 07748, USA
Phone: +1-(732)-420-4578
Fax: +1-(732)-368-8659
Email: gash@att.com

Luyuan Fang
AT&T
Room C2-3B35
200 S.Laurel Avenue
Middletown, NJ 07748
Phone: + 1 732 420 1921
Email: luyuanfang@att.com

Wai Sum Lai
AT&T
Room D5-3D18
200 S. Laurel Avenue
Middletown, NJ 07748
Phone: +1 732 420-3712
Fax:+1 732 368-1919
Email: wlai@att.com

Full Copyright Statement