Yasuhiro Katsube (Toshiba)
Yoshihiro Ohba (Toshiba)
Ken-ichi Nagami (Toshiba)

Two Modes of MPLS Explicit Label Distribution Protocol

<draft-katsube-mpls-two-ldp-00.txt>

Status of this memo

Abstract

This memo discusses characteristics of two types of MPLS protocol operations, which we call 'Edge Control' operation and 'Distributed' operation individually, and proposes that these two mode of protocol operations should be specified as the explicit Label Distribution Protocol for the MPLS.

1. Introduction

Label Switched Routers (LSRs) can forward L3 packets based on fixed length labels, e.g., VPI/VCI field in ATM cells, as well as conventional packet forwarding based on L3 address information. Each LSR, including edge router and possibly host, should exchange control messages with its neighbors in order that they share the common understanding of the relationship between the attached label (or its equivalent) and the specific packet stream.

With regard to the procedure for control message exchange, which are called Label Distribution Protocol, several mechanisms have been proposed[ARISSPEC][FANP][IFMP][TDP]. They are largely classified into two types of operation; one is what we call "Edge Control" operation[ARISSPEC][TDP], and the other is what we call "Distributed" operation[FANP][IFMP].

With regard to the trigger for establishing or releasing LSPs, the Edge Control operations are often understood as Topology-driven approach, while the Distributed operations as Traffic-driven approach. It should be noted that the issue of how the protocol works (either the Edge Control operation or the Distributed operation) is not necessarily coupled with the issue of what the trigger is. Several combinations would be possible instead.

This memo outlines two types of explicit label distribution protocols, discusses characteristics of them individually in terms of the trigger for establishing or releasing LSPs as well as the possible granularity level of the LSPs, and proposes that these two modes of operations should be specified as the explicit Label Distribution Protocol for the MPLS.

2. Two Types of Operation for Explicit Label Distribution Protocol

2.1 Edge Control Operation

(a) Operational overview

In the Edge Control operation, the Label Switched Path (LSP) establishment procedure is initiated by an edge node (an ingress or egress endpoint of the LSP which can be a router or a host) of the MPLS cloud. The initiator transmits MPLS control messages to its neighbor (downstream or upstream depending on the actual procedure) in order to request establishment of the LSP. The control messages convey at least information that specifies stream (e.g., L3 destination address prefix) to be transmitted over the LSP. The information that identifies the label to be used may also be conveyed in the initial message.

The LSR that has received the MPLS initial messages from its neighbor checks the validity of the message, memorizing the notified stream information as well as information to identify the corresponding label (which may be determined by the sender side of the initial messages and notified to the receiver side, or determined by the receiver side and notified to the sender side) at the related interface, and possibly transmitting an acknowledgment to the sender of the initial messages.

After having successfully processed the received MPLS control messages, the LSR reconstructs and transmits the MPLS control messages further downstream or upstream along the path of the stream to request establishment of the LSP. The messages includes the information about the stream determined originally by the edge node (initiator).
The same procedure is performed at every LSR along the path of the stream until the path reaches the node that cannot extend the LSP any more (e.g., another edge node of the MPLS cloud).
An edge-to-edge acknowledgment may be returned to the initiator.

(b) Triggers for the LSP establishment or release

Triggers for the LSP establishment can be, for example, the creation of an L3 forwarding table entry (Topology-driven), or the arrival of any or specific data traffic corresponding to the L3 forwarding table entry (Traffic-driven) at an edge node. Recognition of a group of L3 address prefixes which are reachable through a specific egress edge node can be a trigger for establishment of much aggregated LSPs.

Changes of the paths for existing LSPs in response to L3 route changes would be initiated by the LSR which detect the route change regardless of trigger for the initial establishment. The LSR which detects the route change invalidates the old path by transmitting the control message, which is handled and forwarded hop-by-hop toward the edge node for the existing LSP, and creates the new path by transmitting the control message, which is also handled and forwarded hop-by-hop toward the edge node for the new route.

Triggers to release the LSPs would be deletion of the L3 forwarding entry, or can be decrease of the data traffic activity corresponding to the L3 forwarding table.

(c) Granularity levels of the LSP

Edge routers which initiate the LSP establishment procedure determine the definition of the stream (granularity levels of the LSP) to be transmitted over the LSP. The definition of the stream is included in the establishment message and transmitted hop-by-hop along the path of the stream. A variety of granularity levels can be defined by edge routers, e.g., {dst.prefix}, {BGP next hop}, or {OSPF ABR/ASBR}, depending on the role of the edge node (e.g., just an edge of the MPLS cloud, AS border router, or OSPF Area/AS border router). [ARIS]

Establishment of LSPs with fine granularity such as {src.IP, dst.IP}, {src.IP, multicast group} would also be possible with the Message Passing operation, which would be traffic-driven or request-driven. But, as described in 2.2, LSPs with this fine granularity can also be handled by the Distributed operation.

(d) Other notes

The edge-to-edge massage forwarding in this approach enables to associate several related knowledge with the LSP, e.g., hop-count for the LSP can be notified to edge routers, loop detection or prevention for the LSP becomes possible, and completion of the edge-to-edge LSP can be confirmed by the ingress edge router before transmitting data packets over the LSP.

Processing burden for protocol state management and message handling becomes much larger than the Distributed operation in the case that frequency of establishment, change or release of LSPs are relatively high (e.g., for traffic-driven fine granularity stream, or for IP multicast stream with frequent group membership changes).

2.2 Distributed Operation

(a) Operational overview

In the Distributed Operation, the LSP establishment procedure in an MPLS cloud is initiated by individual LSRs (and edge nodes) in a distributed manner. Each of them transmits MPLS control messages to its neighbor (downstream or upstream depending on the actual procedure) in order to share the mapping relationship between a specific stream and the label dedicated to the stream. The messages will convey at least information that specifies stream to be transmitted with the specific label. The information that identifies the label to be used may also be conveyed by the initial message.

The LSR that has received the MPLS initial messages from its neighbor checks the validity of the message, memorizing the received stream information as well as the corresponding label information (which may be determined by the sender side of the initial MPLS control messages and notified to the receiver side, or determined by the receiver side and notified to the sender side) at the related interface, and possibly transmitting an acknowledgment to the sender of the initial messages.

Unlike the case of the Edge Control operation, exchange of MPLS messages with its neighbor (upstream or downstream) does not trigger exchange of MPLS control messages with its another side of neighbor (upstream or downstream) in an LSR in this case. An MPLS control message exchange for a specific stream between each pair of neighboring LSRs is initiated and carried out independently from the message exchange for the same stream between any other pair of LSRs.

(b) Triggers for the LSP establishment or release

Triggers for the LSP establishment can be, for example, the arrival of any or specific data traffic (Traffic-driven) at individual LSRs and edge nodes on the path. The common rule with regard to the trigger for the LSP establishment (the condition to initiate the LSP establishment) should be configured to all LSRs and edge nodes in the MPLS cloud in order that the LSPs are successfully established in a distributed manner.

The arrival of RSVP Resv messages (Request-driven) at individual LSRs and edge nodes on the path will also be appropriate for the distributed protocol operation. Reception of the Resv message at an LSR from its downstream neighbor triggers the control message exchange with the downstream neighbor (to notify mapping relationship between the stream corresponding to the RSVP flow and the label information to convey the stream), then the LSR transmits the Resv message further upstream. Here we assume the use of current standard RSVP message format with no additional object defined for the MPLS.
The same thing applies to the case of multicast with PIM-SM. Reception of PIM Join messages (Request-driven) at an LSR from its downstream neighbor triggers the control message exchange with the downstream neighbor as well as the LSR transmits the PIM Join message further upstream. Here we assume the use of current standard PIM message format with no additional object defined for the MPLS.

Changes of the paths for existing LSPs in response to L3 route changes are initiated by the LSR which detect the route change regardless of trigger for the initial establishment. The LSR which detects the route change invalidates the mapping relationship between the label and the stream toward its downstream neighbor by exchanging control messages with it, which however does not trigger the control message transmission toward further downstream nodes. The old path will be released by timeout because of, e.g., no data traffic is emitted to the old path or no Path/Resv message transmitted over the old path. Creation of the new path from the LSR that detect the route change will be carried out in the distributed manner similarly to the initial LSP setup procedure.

Triggers to release the LSPs would be, for example, decrease of data traffic activity, or RSVP reservation state expiration at individual LSRs or edge nodes on the path, which keeps principles of distributed operation.

(c) Granularity levels of the LSP

Definition of the stream should be determined by individual LSRs and edge nodes on the path with their own decision since no such information is conveyed hop-by-hop by the control messages in the

Distributed operation. Therefore, the granularity levels provided by the Distributed operation is restricted to the extent that individual LSRs and edge nodes can commonly understand by themselves.

In the case of Traffic-driven setup, LSRs and edge nodes on the path can recognize the stream of L3 level end-to-end granularity individually by referring to the data packets (e.g., {src.IP, dst.IP} and {src.IP, multicast group}). In addition, they can recognize the stream of {src.IP, dst.prefix} granularity individually when they are guaranteed to have the forwarding table entry with the same aggregation level given by the routing protocol or by configuration.

In the case of Request-driven setup, each of the LSRs can recognize the stream with application to application granularity by referring to the RSVP Resv messages (e.g., {src.IP/port, dst.IP/port}), or recognize the stream with L3 level end-to-end granularity by referring to the PIM Join messages (e.g., {RP, multicast group} or {src.IP, multicast group}).

As the data packets, the RSVP Resv messages, and the PIM Join messages travel along the path of the LSP with the information of those stream definition, they perform almost the same role as the edge-to-edge

Edge Control case, which facilitate the LSP control with the Distributed operation.

(d) Other Notes

Although no information with edge-to-edge importance can be shared through the Distributed operation, overall procedure is simple and it is easy to follow dynamic changes in router state, e.g., unicast routing, multicast group membership, or RSVP reservation state. Various knowledge related to the LSP such as hop-count, existence of loop cannot be obtained in the Distributed operation.

3. Desirable Protocol Operations for Individual Types of LSPs

3.1 Unicast LSP

3.1.1 Unicast LSP with Arbitrary Granularity Level

When the MPLS cloud should provide LSPs for aggregated streams with various granularity levels, the use of Edge Control operation is desirable. The granularity level should be determined by edge nodes (either ingress or egress), then should be notified by MPLS control messages hop-by-hop to all LSRs on the path of the stream.

The LSP establishment can be triggered by creation of forwarding table entries (Topology-driven) or the arrival of traffic corresponding

to the table entry (Traffic-driven). The release of the LSP can be triggered by the deletion of the forwarding table entries or can be triggered by the decrease of traffic activities corresponding to the table entry.

Figure 1 shows examples of a message sequence for unicast LSP setup with the Edge Control operation. The sequences initiated by an ingress edge (like TDP) and the sequence initiated by an egress edge (like ARIS) are shown. Note that the detailed procedure should be specified by the MPLS WG.

```
    Ingress======== LSR1 ======== LSR2 ======== LSR3 ========Egress

      TRG   req           req           req           req
```

```
         |-------->++++|-------->++++|-------->++++|-------->++
            ack           ack           ack           ack     |
         <--------|++++<--------|++++<--------|++++<--------|++

                (i) Ingress Initiated Sequence


    Ingress======== LSR1 ======== LSR2 ======== LSR3 ========Egress

            req            req           req           req   TRG
        +<---------|++++<--------|++++<--------|++++<---------|
        |   ack         |   ack         |   ack         |   ack
        +---------->   +--------->   +--------->    +--------->

                (ii) Egress Initiated Sequence

     (TRG = "creation of forwarding entry"
             or "arrival of data packets")

    Fig.1  Examples of Message Sequence for Arbitrary Granularity
```

3.1.2 Unicast LSP with Limited Granularity Level

When the MPLS cloud provides unicast LSPs for specific end-to-end L3 streams on-demand (Traffic-driven), it can adopt the Edge Control operation since the end-to-end L3 stream (specified by {src.IP, dst.IP}) is just one of the granularity levels described in 3.1.1. But it should be noted that provision of traffic-driven LSPs for end-to-end L3 streams requires much frequent establishments or releases of LSPs compared with aggregated LSPs. Distributed operation which is more lightweight than Edge Control operation

may be preferable in this case. As described in 2.2, it is possible to provide, for example, {src.IP, dst.prefix} level granularity in a domain whose routers share the forwarding entry with the same level of network mask.

Figure 2 shows an example of the message sequence for unicast LSP setup with the Distributed operation triggered by data traffic. Note that the detailed procedure should be specified by the MPLS WG.

```
Ingress======== LSR1 ======== LSR2 ======== LSR3 ========Egress
      packet         packet        packet        packet
->  - - - - - - ->   - - - - - - ->  - - - - - - ->  - - - - - - ->
   TRG   req      TRG  req      TRG  req      TRG  req
     |-------->+   |-------->+   |-------->+   |-------->+
        ack   |       ack   |       ack   |       ack   |
     <---------+   <---------+   <---------+   <---------+

 (TRG = "arrival of data packets")
```

Fig.2 Example of Message Sequence for Fine Granularity

3.2 Multicast LSP

When the MPLS cloud provides LSPs along source-based or shared multicast trees, point-to-multipoint LSPs will be established whose origination points are either the ingress edge node closest to the source or the RP for the PIM-SM.

The traffic-driven, Distributed operation is straightforward in the case of dense mode protocol such as DVMRP in terms of the initial setup procedure as well as addition or deletion of group members. Triggered by the arrival of multicast packets, each LSR can establish dedicated labels to its downstream neighbors using the Distributed operation.

The request-driven, Distributed operation is straightforward in the case of sparse mode protocol such as PIM-SM in terms of initial setup as well as addition or deletion of group members. Triggered by the arrival of PIM Join messages from the downstream neighbors, each LSR can establish dedicated labels to its downstream neighbors using the Distributed operation. Note that inclusion of label information in the PIM Join message may be enough for label establishment in some cases as described in [TAG]. But in the case that the label value is changed between neighboring LSRs as described in [KATSU], inclusion

of label information in the PIM Join message alone is not enough but additional message handshake between neighboring LSRs is necessary.

Figure 3 and 4 show examples of the message sequence for multicast LSP setup with the Distributed operation in the traffic-driven case and request-driven case individually. Note that the detailed procedure should be specified by the MPLS WG.

```
             ack            ack          ack
          <---------+   <---------+  <----------+
             req    |       req   |      req    |
          |-------->+   |-------->+  |--------->+
         TRG           TRG          TRG
           packet         packet       packet
     ->  - - - - - ->  - - - - - ->  - - - - - ->
     Ingress======== LSR1 ======== LSR2 ========Egress
                       |       packet
                       | - - - - - - - - - - - ->
                       +========================Egress
                                  req
                        |----------------------->+
                                  ack            |
                        <-----------------------+
```

(TRG = "arrival of data packets")

Fig.3 Example of Message Sequence for Multicast LSPs (Traffic-driven)

```
             ack            ack          ack
          <---------+   <---------+  <----------+
             req    |       req   |      req    |
          |-------->+   |-------->+  |--------->+
         TRG          TRG           TRG
           PIM Join      PIM Join      PIM Join
          <- - - - -    <- - - - -    <- - - - -
     Ingress======== LSR1 ======== LSR2 ========Egress
                       |
                       +========================Egress
                                PIM Join
                        <- - - - - - - - - - - -
                       TRG       req
                        |----------------------->+
                                 ack             |
                        <-----------------------+
```

(TRG = "arrival of PIM Join")

Fig.4 Example of Message Sequence for Multicast LSPs (Request-driven)

3.3   Unicast/Multicast LSP with RSVP

With regard to the LSP establishment in response to the creation of RSVP reservation state (Request-driven), the Edge Control operation initiated by edge nodes does not adequate since each LSR must forward RSVP Resv messages upstream after it succeeds to establish labels toward its downstream neighbors, which requires distributed LSP control operation rather than operation initiated by edge routers.

The procedure will almost the same as the case with PIM-SM. Triggered by the arrival of RSVP Resv messages from the downstream neighbors, each LSR would establish dedicated labels to its downstream neighbors using the Distributed operation. Note that the inclusion of label information in the RSVP Resv message may be enough for label establishment in some cases as described in [DAVIE][VISWA]. But in the case that the label value is changed between neighboring LSRs as described in [KATSU], inclusion of label information in the RSVP Resv message alone is not enough but additional message handshake between neighboring LSRs is necessary.

Figure 5 shows an example of the message sequence for rsvp-driven LSP setup with the Distributed operation. Note that the detailed procedure should be specified by the MPLS WG.

```
          ack            ack          ack
       <---------+   <---------+  <----------+
          req    |       req   |      req    |
       |-------->+   |-------->+  |--------->+
      TRG          TRG           TRG
          Resv          Resv          Resv
       <- - - - -    <- - - - -    <- - - - -
  Ingress======== LSR1 ======== LSR2 ========Egress
                    |
                    +========================Egress
                              Resv
                     <- - - - - - - - - - - -
                    TRG       req
                     |----------------------->+
                              ack             |
                     <------------------------+
```

(TRG = "arrival of RSVP Resv message")

Fig.5 Example of Message Sequence for LSPs with RSVP
(Request-driven)

4. Security Consideration

Security issues are not discussed in this memo.

5. Conclusion

Based on to the discussion above, we propose that the two mode of explicit label distribution protocols, which we call "Massage Passing" operation and "Distributed" operation, should be supported.
Either of them would be utilized according to the stream granularity trigger, and configuration (p-p/p-mp) of the LSP.

6. References

[ARIS] A.Viswanathan, N.Feldman, R.Biovie, and R. Woundy, "ARIS: Aggregated Route-Based IP Switching", draft-viswanathan-aris-overview-00.txt, March 1997.

[ARISSPEC] N. Feldman, A. Viswanathan, "ARIS Specification", draft-feldman-aris-spec-00.txt, March 1997.

[DAVIE] B. Davie, Yakov Rekhter, and Eric Rosen, "Use of Label Switching With RSVP", draft-davie-mpls-rsvp-00.txt, May 1997.

[FANP] K. Nagami, Y.Katsube, Y. Shobatake, A. Mogi, S. Matsuzawa, T. Jinmei, and H. Esaki, "Toshiba's Flow Attribute Notification Protocol (FANP) Specification", RFC2129, April 1997.

[IFMP] P. Newman, W. L. Edwards, R. Hinden, E. Hoffman, F. C. Liaw, T. Lyon, and G. Minshall, "Ipsilon Flow Management Protocol Specification for IPv4", RFC1953, May 1996.

[TAG] Y. Rekhter, B. Davie, D. Katz, E. Rosen, G. Swallow, and D. Farinacci, "Tag Switching Architecture - Overview", draft-rekhter-tagswitch-arch-00.txt, Jan. 1997.

[TDP] P.Doolan, B.Davie, D.Katz, Y.Rekhter, and E.Rosen, "Tag Distribution Protocol", draft-doolan-tdp-spec-01.txt, May 1997.

[VISWA] A. Viswanathan and V. Srinivasan, "Soft State Switching - A Proposal to Extend RSVP for Switching RSVP Flows -", draft-viswanathan-aris-rsvp-00.txt, March 1997.

7. Authors Addresses

Yasuhiro Katsube
R&D Center, Toshiba Corporation,
1 Komukai-Toshiba-cho, Saiwai-ku,
Kawasaki, 210, Japan
Email: katsube@isl.rdc.toshiba.co.jp

Yoshihiro Ohba
R&D Center, Toshiba Corporation,
1 Komukai-Toshiba-cho, Saiwai-ku,
Kawasaki, 210, Japan
Email: ohba@csl.rdc.toshiba.co.jp

Ken-ichi Nagami
R&D Center, Toshiba Corporation,
1 Komukai-Toshiba-cho, Saiwai-ku,
Kawasaki, 210, Japan
Email: nagami@isl.rdc.toshiba.co.jp