

Multi-Protocol Label Switching (MPLS)

Overview and Applicability

Author : Arup Acharya (arup@us.ibm.com)

Group : Mandis Beigi, Raymond Jennings, Reiner Sailer,
Dinesh Verma (Manager)

Benefits of MPLS

- § MPLS provides the ability to forward packets over arbitrary non-shortest paths, i.e. it provides a circuit switching service in a hop-by-hop routed network.
- § For non-IP based networks such as ATM or frame relay, it provides a IP based control plane (routing, path selection, reservation) instead of technology specific control protocols (e.g. PNNI). MPLS thus provides a unifying control architecture for both connectionless and connection-oriented switching/routing hardware.
- § It provides a mechanism to group a related set of packets together by assigning a common "label" and isolating one group of packets from another. Thus, a label-switched path (LSP) can be setup to provide a generic tunneling service, e.g.
 - connect segments of a VPN over a public network,
 - interconnect two non-IP based networks (instead of say L2TP), or
 - associate a common forwarding rule for packets sharing the same label, e.g. class of service.

LSPs can be nested through the use of a label stack. LSPs can also be concatenated. MPLS provides for both point-to-multipoint and multipoint-to-point LSPs. The former is used for multicasting while the latter is used to aggregate traffic from multiple entry points onto a common exit point.

- § Labels have been defined for most layer2 technologies (ethernet, PPP, ATM, frame relay) and as a result, MPLS services can be offered over a collection of heterogeneous networks.
- § The original motivation for MPLS was to enable fast switching, by replacing route lookup for a variable length IP destination address, with an exact match of a fixed, predefined number of bits. However, with the advent of fast route lookup algorithms and routing hardware, usefulness of MPLS in this regard is limited. Nevertheless, the use of labels to explicitly identify a common group of packets rather than matching variable parts of the packet header, may be useful in other contexts that require quick indexing into a table of rules. For example, packets that receive a common security inspection may be identified with a common label. Or, in load balancers for web-servers, connections that belong to a common session may be assigned a common label so that packets for that session are routed to the same server.¹

¹ These should be considered as "nonstandard" usage of MPLS labels.

§ Since the interpretation of labels is independent of the control protocols, new protocols can easily be supported. The switching hardware typically supports the following operations:

- link a label with a packet scheduling behaviour : this is currently under discussion where each label should represent a single behaviour or whether specific bits in the label header encode specific behaviours.

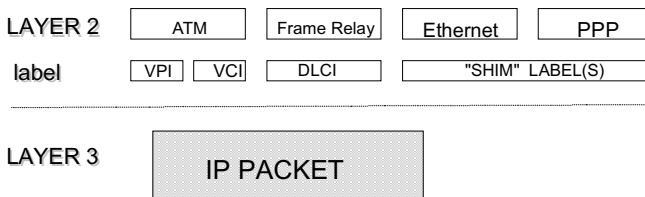
Labels

Label definition

A label is a fixed width identifier to group a related set of packets. Packets that share a common attribute, i.e. belong to a “forwarding equivalence class” (FEC), are assigned a common label. For example, all packets that are routed to a common exit router in a service provider’s network are assigned the same label.

Format of a label

MPLS supports three different types of labels: on ATM hardware, it uses the VPI/VCI fields on each cell; on frame relay hardware, it uses the Data Link Connection Identifier (DLCI) field in each frame as the label;; everywhere else, MPLS uses a new, generic shim label, which wedges



between layers 2 and 3. This label is 20 bits wide, with 3 experimental bits (for QoS/Diffserv), a 8 bit TTL field (to prevent loops) and a S bit (to enable label stacking). The S bit at the bottom of a label stack is set to 1; the S bit on all other labels is set to 0.

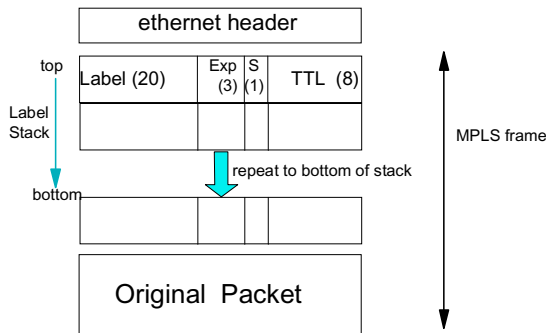
Label Stacking

Label stacking enables creation of tunnels. For example, a network service provider may create a tunnel between a given pair of provider edge routers. This tunnel may multiplex traffic for different VPNs. In this case, a two-level label stack is used , where the outer label identifies traffic

that is carried on a common LSP between the edge routers. The inner label identifies a specific VPN.

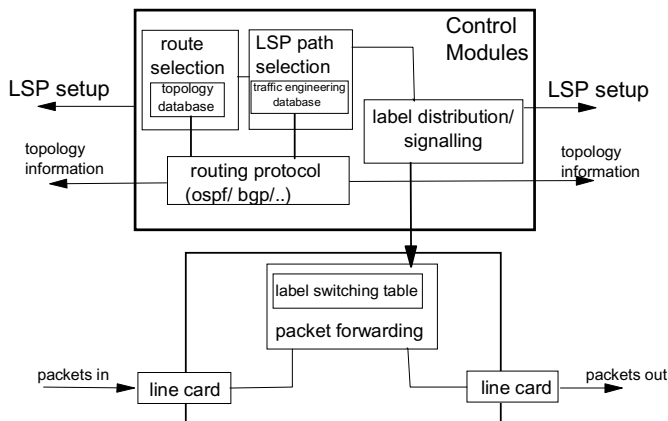
Label switching router

A label switching router decouples control from packet forwarding. Packet forwarding is typically done in hardware. A set of entries of the form <input port, input label, output port(s), output



label(s)> govern the packet forwarding decisions.

The (label) forwarding tables are populated by a control module, which is responsible for distributing and assigning labels. The control module could be a unicast routing protocol together with a label distribution protocol (LDP) that distributes labels to mirror shortest path routes. A second instance of a control module is a traffic engineering module that enables paths within a network to be specified explicitly : RSVP-TE and CR-LDP are two proposals to reserve

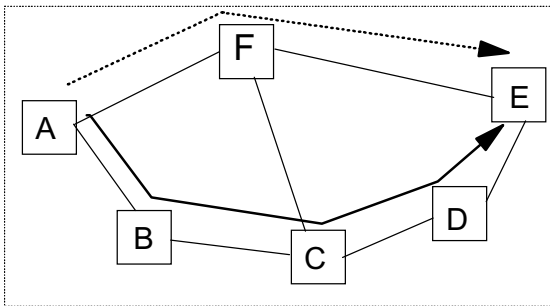


resources and assign labels along a explicitly specified path. Another instance of a control module is a VPN module on a network service provider's edge router, that creates VPN specific label forwarding tables.

Virtual Private Networks (VPNs)

Since MPLS allows different modules to assign labels to packets using a variety of criteria, it decouples the forwarding of a packet from the contents of the packet's IP header.

A unicast routing module that distributes labels (LDP) that mirror shortest path routes as calculated by a standard unicast routing protocol such as OSPF. As seen below, this would create a LSP from edge router A to E mirroring the minimum hop paths A -- F --- E.



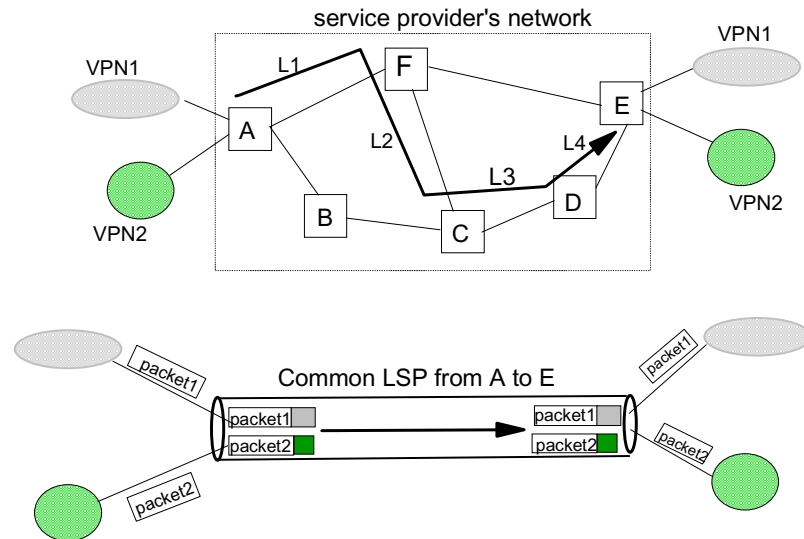
A traffic engineering module that sets up explicit routes through either CR-LDP or RSVP-TE. For example, the LSP in this case from A to E could consist of A--B--C--D--E.²

A VPN module that builds VPN-specific routing tables using BGP and distributes routes to interconnect different VPN segments.

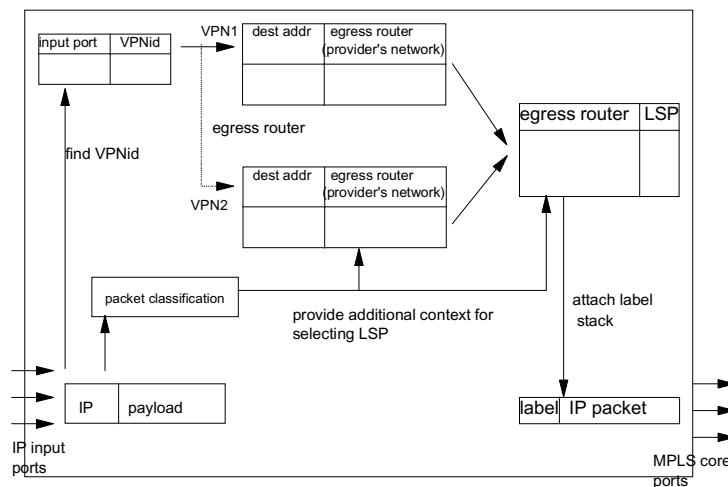
The following VPN illustration shows a service provider that uses a single shared LSP to serve two customer VPNs. The LSP, as implied by the sequence of labels, {L1, L2, L3, L4, L5}, acts as a tunnel for all packets originating at the edge router A and terminating at the edge router E. At E, packets exiting from the tunnel, need to be forwarded to the appropriate access link (leading to either VPN1 or VPN2). This is achieved by a two-level label stack : packets intended for VPN1 are labeled with a “blue” label at A, while packets for VPN2 are labeled with a “green” label. This label serves as the inner label. Prior to forwarding a packet into the network, A inserts L1 as the outer label : this forces the packet to be switched to the edge router B using the LSP (A,F,C,D,E). At E, the forwarding action associated with label L4 is a *stack pop* , i.e. the outer label L4 is removed and the next label (blue or green) determines whether the packet is forwarded to VPN1 or VPN2.

² This set can be considered as a policy input.

This example assumes that MPLS operates only within the provider's network. In an alternative proposal, it is possible for the customer's edge routers (e.g. those connected directly to A and E) to participate in MPLS.



When a VPN service is offered by a provider through MPLS, its edge routers need to maintain an additional set of routing/label tables.



Each supported VPN requires a set of routing tables for addresses (public/private) reachable within the VPN³. Typically, each access port connects to a single VPN (segment) and the incoming port is used to determine the VPNid. The destination address along with additional information from the packet header (such as TOS, port numbers) guides selection of the appropriate LSP : note that depending on how the EXP bits in the label are used, all packets to a

³ BGP extensions have been proposed to propagate reachability of VPN segments amongst each other, and thus populate VPN specific routing tables at edge routers.

common egress edge router may either be mapped to a single LSP (with appropriate setting of the EXP bits to denote a class of service) or each class of service may be assigned a different LSP.

The diagram above assumes that all VPN connectivity between a given pair of edge routers use a shared LSP in the provider's network. This is implied through the use of a common table for mapping edge routers to LSPs. Alternatively, it is possible to provide separate LSPs for per-VPN connectivity at the expense of higher label usage within the provider's network. This would require a separate per-VPN table for mapping egress routers to LSPs.

MPLS and Differentiated Services (DiffServ)

DiffServ is a framework for providing Internet QoS in a scalable manner. DiffServ involves marking packets with a tag indicating the requested QoS on a per hop basis, as opposed to circuit-oriented per-flow reservation (as in Integrated Services/RSVP). For this purpose, the IP precedence and Type-of-service (TOS) bits in a IP packet header as been redefined for use as DiffServ CodePoints (DSCP). The DSCP byte in the packet header determines the per-hop behaviour (PHB). Currently, there are four proposed mappings of DSCP :

- Best effort
- Expedited forwarding
- Assured forwarding
- Class selector (backward compatibility with IP precedence)

There are two broad methods for mapping DiffServ classes onto MPLS label-switches paths : L-LSPs and E-LSPs. L-LSP or label inferred LSP associate a layer 3 Diffserv codepoint with a specific LSP, i.e. one LSP is dedicated to each DiffServ codepoint. This could potentially tie up network resources unnecessarily, e.g. DSCP has enough bits to provide 64 codepoints, and with this method, 64 separate LSPs will be needed. E-LSPs, on the other hand, exploit the 3 experimental bits (EXP) in the MPLS label to map upto 8 DSCPs. Thus, a single E-LSP can be used to carry traffic corresponding to different DSCPs. The mapping of DSCP to the EXP bits is done by ingress label switching router. At each hop of the E-LSP, the EXP bits in the label determine the per-hop behaviour received by packets forwarded on that LSP.

Policy controls

Following are some key policy controls needed for a MPLS network:

- the set of LSPs to provide a desired connectivity between edge routers in a traffic engineered network : this would trigger RSVP-TE/CR-LDP signaling to setup LSPs. (traffic engineering policy)
- definition of FECs and their mappings to LSPs : this would populate the label information base at the edge routers, mapping some combination of IP header fields and incoming port to outgoing label/ports.

--- for VPN connectivity, the list of segments (edge routers) to be included in the VPN can be specified as a policy input.

Current usage of MPLS

MPLS today is primarily targeted for use by network service providers. MPLS allows service providers to offer a VPN service with QoS (DiffServ), including intranets and extranets. Since traffic from multiple customers can be multiplexed over a common set of LSPs, MPLS provides a scalable method for service providers to meet customers' connectivity needs. For customers of a MPLS based VPN service, a key advantage is simplification of routing and management overheads : there is no coupling between the customer's routing tables and that of the service provider. The provider's network appears as a private IP backbone connecting different sites of a customer's organisation and the customer typically will use the service provider's network as the default route to reach all other sites in the customer's organisation. Additionally, for service providers, the traffic engineering features offered by MPLS (to setup explicitly routed non-shortest paths) are useful to better manage traffic and link utilization in the provider's network. For providers that use an ATM network, MPLS offers a IP-based routing and signaling mechanism to provide direct IP services, instead of an IP overlay over ATM signaling and routing.

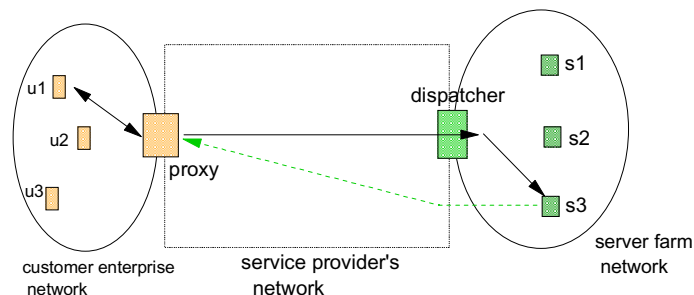
While MPLS is directed towards carriers/ISPs, its benefits are also applicable to large enterprise networks and it may become the defacto method by which service providers offer VPN services to multi-site enterprise networks.

Possible future usage of MPLS

Server farm networks and maintaining affinity between users and servers :

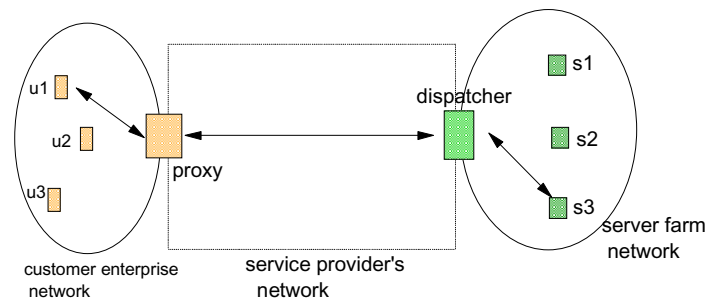
A typical architecture for web access today consists of a proxy/cache in a customer's enterprise network, and a web farm interconnected through a service provider's network. Instead of a single TCP/HTTP connection between a user and a web server, there exists two HTTP connections : one between the end-user and the proxy, and a second connection from the proxy to a web server. One of the commonly used methods advertises a single IP address for all the servers, and a network dispatcher routes incoming connection requests to a specific server typically, by using link-layer addresses within the server farm network.

This has two undesirable consequences. Firstly, user connections that belong to a common session may be routed to different servers by the dispatcher since the source IP address in the



connection request is that of the proxy. Alternatively, the dispatcher may route all requests from a given proxy to a single server (since the source address is the same for all such requests). This may lead to uneven loading of the individual servers. Second, the dispatcher in some implementations, may terminate the incoming TCP connection, inspect the HTTP request and then based on the HTTP request, setup another TCP/HTTP connection to a specific server. In effect, the dispatcher is forced to implement content-based routing of incoming requests.

MPLS labels may offer a method to avoid the above shortcomings. Assuming that the service provider's network is MPLS enabled, we propose that a second (inner) label be used to identify the association between users and servers. Using the label stacking feature of MPLS, the outer label is used for routing within the service provider's network (from the proxy to the dispatcher), while the inner label is used by the dispatcher to route an incoming request to a specific server.



Note that this does not require the connection to be terminated at the dispatcher. The proxy maintains a mapping of user requests to labels, and thus succeeding requests from the same user (modulo a timeout period) are associated with the same inner label. Further, since the proxy has multiple labels at its disposal, it will use different labels to disambiguate multiple users. This ensures that requests from different users behind a common proxy are not routed by the dispatcher to a single server. In effect, with this method, the responsibility of load-balancing user requests is now shared between the proxy and the dispatcher. The dispatcher controls macro-level load balancing across proxies (by controlling the labels distributed) while individual proxies have the responsibility of balancing local user requests across labels it has received.

A second application of MPLS could be to implement the server farm network : there is a need to route requests within the server farm network either using link-layer addresses (since all servers are identified externally with a common IP address) or by encapsulating the incoming packets with an outer IP header containing the actual IP address of the destination server. The advantage of using MPLS in the server farm network is that labels can be used to route packets from the dispatcher to the appropriate server. Coupled with the label stacking method discussed above, the outer label is used to route within the service provider's network while the inner label is used to route/switch within the server farm network⁴.

Voice over IP :

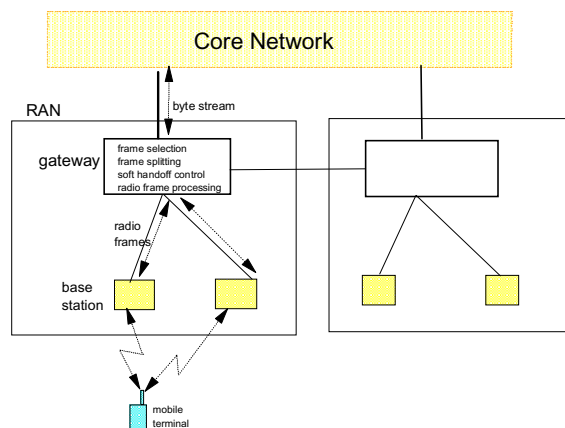
VoIP packets are typically small and the UDP/IP header represents a significant overhead for such packets. More importantly, the network must support stronger QoS guarantees in terms of delay, jitter and packet loss in order for VoIP to offer comparable service to regular telephony. MPLS may offer solutions to both these issues within the enterprise networks. First, the MPLS

⁴ This section benefited greatly from discussions with Anees Shaikh and Renu Tiwari.

label (instead of a full IP/UDP header) may be used to transport voice packets thereby reducing the bandwidth usage. Secondly, label switched paths along with DiffServ / IntServ allows voice calls to be accorded QoS similar to that of a circuit-switched connection. A possible architecture consists of per-connections LSPs within the enterprise networks, which are grouped into “trunk” LSPs within the core.

Radio Access Networks in 3rd Generation Wireless networks :

The goal of 3rd generation (3G) wireless networks is to enable high speed data services to mobile wireless devices, in addition to voice. A 3G system will consist of a IP based core network, a radio access network (RAN) and (wideband)CDMA wireless cells. A CDMA RAN consists of multiple base-stations that offer a wireless interface to mobile terminals and mobile terminals may simultaneously be in contact with multiple of these base-stations. For traffic to the mobile terminal, the RAN gateway processes incoming byte stream into short “radio frames” (about 20 bytes), and replicates each radio frame among the base-stations currently in contact with a mobile terminal (“frame splitting”). On the reverse channel, the gateway collects copies of a transmitted frame (possibly, with errors) from multiple base-stations and combines them into a single error-free byte stream. The RAN gateway is also responsible for power control at the base-stations. Due to terminal mobility, the set of base-stations in contact with the terminal changes leading to a “soft handoff”. If this set spans multiple RANs, then the frames are forwarded between the RAN gateways for frame splitting/combining. Currently, a “hard handoff”



is invoked when the RAN gateway changes due to terminal mobility.

There is ongoing discussion in different standards organizations (3GPP2, MWIF, 3GIP) on how to replace legacy mobile telephony signaling (such as IS-41/MAP) with IP based solutions (e.g. Mobile IP) within the core network. It is also likely that MPLS will be used as the underlying transport within the core network.

Recently, there has also been a strong interest to extend an all-IP solution to the radio-access network, which has a different set of constraints than the core network. These constraints include very stringent delay and jitter requirements on how the frames are transported between the RAN gateway and base-stations, and also the time to complete a soft handoff. Power control amongst the set of base-stations in contact with a mobile terminal is an important requirement. The gateway and base-stations are connected by dedicated point-to-point links. Mobility within a RAN is handled by local mechanisms that are transparent to wide area mobility protocols.

The goal of an IP based RAN is to create a switched peer-to-peer architecture spanning multiple base-stations and place more intelligence (e.g. Mobility support, power control) at the base-stations, rather than relying on a centralized RAN gateway for control. MPLS could become an important component of an IP based RAN architecture. Given the short size of radio frames, it would be inefficient (especially on the wireless links) to use full IP headers and hence, an IP routed network to interconnect the base-stations. Thus, a label switched network could provide an efficient transport network both within the wired and wireless RAN segments. DiffServ could be used to ensure that the delay and jitter constraints are satisfied. Newer proposals for mobile IP optimized for local mobility may offer a solution for mobility control within the RAN, that is harmonized with wide-area mobility support (instead of being transparent). Additionally, the radio related control mechanisms will need to be recast within a IP based signaling and control framework.

Optical Networks :

MPLS is currently being considered for use as the control plane for controlling a network of optical switches. An optical network consists of optical cross-connects (OXC) interconnected by optical links. Each optical link carries some number of wavelengths, and a OXC switches between a wavelength from an incoming link to an outgoing link. A lightpath consists of a sequence of OXCs, with each OXC converting wavelengths between its input and output links on the path. In order to set the cross-connect table entries along a lightpath, an appropriate signaling mechanism is required. The signaling mechanisms offered by MPLS such as RSVP-TE have been proposed for this purpose, with modifications to carry additional parameters specific to optical networks (e.g. bi-directional LSPs, protection support). Each wavelength (or a sub-channel) effectively represents a label, and the sequence of wavelengths in a lightpath defines a label-switched path in this context.