



Introduction to MPLS

Peter J. Welcher

Introduction

A note from my employer: Chesapeake Network Solutions has changed names to Mentor Technologies. The intent is to clarify our identity for our partners, customers, and investors. The revised URL for my previous articles is shown at the bottom of this page. (Yes, they still are there!)

I've been looking at MPLS, **M**ultiprotocol **L**abel **S**witching, over the last couple of months. Cisco has some pretty good slideware on the topic, but there have been some things I had wanted to drill down on and understand better. I can't say for sure that I'm all the way there yet. But this month I'd like to share with you some of the basics of MPLS. It may take another article or two to finish this topic.

The IETF working group information (and list of related draft documents) for MPLS can be found at:

- <http://www.ietf.org/html.charters/mpls-charter.html>

For an overview of MPLS, see also:

- <http://www.ietf.org/internet-drafts/draft-ietf-mpls-framework-05.txt>
- <http://www.ietf.org/internet-drafts/draft-ietf-mpls-arch-06.txt>

I found both of these fairly readable. But then, I've seen enough Cisco slideware that I have in my head some basic pictures and ideas of MPLS.

What is MPLS?

MPLS stands for Multiprotocol Label Switching. Multiprotocol because it might be applied with any Layer 3 network protocol, although almost all of the interest is in using MPLS with IP traffic. But that doesn't actually give us any idea what MPLS does for you (we'll get to that momentarily).

Depending on which vendor you ask, MPLS is the solution to any problem they might conceivably have. So the question "What is MPLS" could have a **lot** of right answers. The presentations from this Spring's MPLS Forum were all over the place on precisely this.

For me, MPLS is about gluing connectionless IP to connection-oriented networks. Six months ago, I'd have said "gluing IP to ATM", but now there's a big push on to use MPLS to mate IP to optical networks. The IETF draft documents refer to this as the "shim layer", the idea that MPLS is something between Layer 2 and Layer 3 that makes them fit better (and perhaps carries a small amount of information to help that better fit).

MPLS started out as Tag Switching. Ipsilon (remember them?) was the company that got the MPLS buzz started. Back then, there were perhaps two key insights. One was that there is no reason an ATM switch can't have a router inside it (or a router have ATM switch functionality inside it). Another was that once you've got a router on top of your ATM switch, you can use dynamic IP routing to trigger virtual circuit (VC) or path setup. In other words, instead of using management software, or human configuration, or (gasp!) even ATM routing (PNNI) to drive circuit setup, dynamic IP routing might actually drive the creation of circuits. You might even have a variety of protocols for different purposes, each driving Label Switch Path establishment.

I've been thinking of this as avoiding the hop-by-hop decision making, by setting up a "Layer 2 fast path" using tags (think ATM or Frame Relay addressing) to move things quickly along a pre-established path, without such "deep analysis". The packet then needs to be examined closely exactly once, at entry to the MPLS network. After that, it is somewhere along the path, and forwarding is based on the simple tagging scheme, not on more complex and variable IP headers. The U.S. postal system seems to work like that: forward mail to a regional center, do handwriting recognition once, apply some sort of infrared or ultraviolet bar code to the bottom edge of the envelope, from there onwards, just use the bar code to route the letter. When you start thinking about fast forwarding with Class of Service (CoS), then incoming interface, source address, port and application information, all might play a role in the forwarding decision. By rolling the results into one label per path the packet might take, subsequent devices do not need to make such complex decisions.

Fairly soon after the basic idea of Tag Switching got publicized, Cisco got visibly involved, and then so did all the other vendors of course. For a couple of years now, Cisco Tag Switching in the 7000 series has allowed using Tag Switching on high-speed IP networks. This is migrating right now, to support the final standardized Label Switching. Other Cisco platforms now supporting MPLS: LS1010, 3600 (the release notes for 12.1(3) T say 2600), 12000 GSR series.

It now looks like optical networking devices will be capable of fast circuit establishment. Lucent has announced an "Optical Router", using 256 very small mirrors on a chip, steered under electrical control. Agilent (HP) and Texas Instruments have announced liquid or gel-based chips where current turns the fluid to a reflective surface, deflecting light from one waveguide into another. For me, all these devices deserve a title like Optical DACS (Cross-Connect Switch), but who asked me? (The Cisco press release for the Monterey Networks acquisition refers to their optical cross-connect technology). These devices are **not** routers in the sense of looking into packets and determining path dynamically. They are routers in the sense of figuring out and plumbing a path through multiple Layer 1 devices. I prefer not to call that routing.

MPLS ties to optical by using the idea that when a route to a specific destination or group of destinations is propagated, a light path might also be set up. This light path could then be used by packets going to that destination or group of destinations, getting them there faster (one hopes) than if every router or device along the path examined the Layer 3 header. Actually having a program examine the Layer 3 information would involve converting the light to and from electrical signals at each step along the way.

So we have several media where MPLS is being considered:

- high-speed IP backbones
- legacy ATM
- MPLS-capable ATM
- optical

Frame Relay MPLS is also receiving some consideration by vendors other than Cisco.

You might be wondering if anyone is actually doing MPLS, or is this cutting-edge stuff? Well, AT&T is offering a service called IP Frame Relay, with IP access via Frame Relay to an MPLS network. This uses the Cisco-driven idea of MPLS-based VPN's, discussed very lucidly in RFC 2547. MCI has announced a similar service, called Business Class IP (which I can't find on their service offering web pages). See also:

- AT&T IPFR: <http://www.ipservices.att.com/data/framerelay/framer.html>
- AT&T IPFR: http://www.ipservices.att.com/databriefs/profiles/31_earthtech.html
- MCI Business Class IP: http://www.nwfusion.com/archive/2000/96405_05-15-2000.html
- MCI Business Class IP: <http://www.nwfusion.com/news/2000/0510mciprivate.html>
- MCI Business Class IP: <http://nwfusion.com/columnists/2000/0529rohde.html>

Looking More Deeply Into MPLS

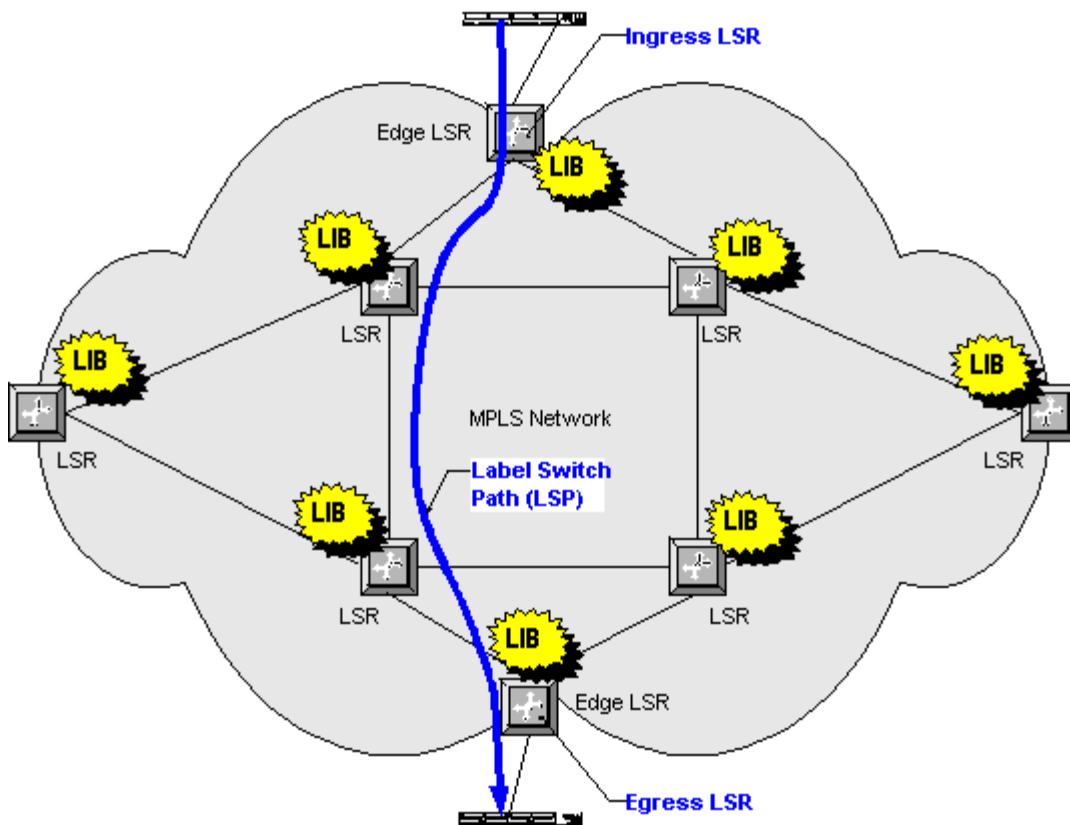
A router supporting MPLS is a Label Switch Router, or LSR. An edge node is an LSR connecting to a non-LSR. An ingress LSR is the one by which a packet enters the MPLS network, an egress LSR is one by which a packet leaves the MPLS network.

Labels are small identifiers placed in the traffic. They are inserted by the ingress LSR, and ultimately removed by the egress LSR (so nothing will remain to perplex the non-MPLS devices outside the MPLS network). For IP-based MPLS, some bytes are inserted prior to the IP header. For ATM, the VPI/VCI addressing is the label. For Frame Relay, the DLCI is the label. For optical, I imagine the label is the optical fiber and/or wavelength being used (implicit label), perhaps combined with some actual label. To read more about MPLS labels with LAN and PPP, see:

- <http://www.ietf.org/internet-drafts/draft-ietf-mpls-label-encaps-07.txt>

As traffic transits the MPLS network, label tables are consulted in each MPLS device. These are known as the Label Information Base, or LIB.

By looking up the inbound interface and label in the LIB, the outbound interface and label are determined. The LSR can then substitute the outbound label for the incoming, and forward the frame. This is analogous to if not exactly the way Frame Relay and ATM behave as they send traffic through a virtual circuit. For that matter, IBM High Performance Routing, HPR, behaves similarly as far as how it actually forwards data. The labels are locally significant only, meaning that the label is only useful and relevant on a single link, between adjacent LSRs. The adjacent LSR label tables however should end up forming a path through some or all of the MPLS network, a Label Switch Path (LSP), so that when a label is applied, traffic transits multiple LSRs. If traffic is found to have no label (only possible in an IP MPLS network, not Frame Relay or ATM), a routing lookup is done, and possibly a new label applied.



As you read about MPLS you'll encounter "Forwarding Equivalency Class", or FEC. This just refers to the idea that all sorts of different packets might need to be forwarded to the same next hop or along the same MPLS path. The FEC is all the packets to which a specific label is being applied. This might be all packets bound for the same egress LSR. For a Service Provider, all packets with a given Class of Service (CoS) bound for a certain AS boundary router, or matching certain CIDR prefixes. For a large company, all packets matching certain route summaries.

Traffic may actually bear multiple labels, a stack of labels. Only the outermost (last) label is used for forwarding. The label table in a LSR may cause the outermost label to be removed. This is called a "label POP". This is useful for MPLS Tunneling, which is useful for Traffic Engineering.

Binding is the process of assigning labels to FECs. A Label Distribution Protocol (LDP) is how MPLS nodes communicate and bind labels. Think of an LDP as being an official way for one LSR to say to another "let's use this label to get stuff to this destination really fast". More than one LDP is being contemplated, each specifically designed for a purpose. However, LDP without further qualification refers to the standard LDP for setting up Label Switched Paths in response to IP routing. After LDP has run hop-by-hop, the MPLS network should have paths from ingress to egress LSR, along which only labels are used. Such paths are called Label Switch Paths (LSPs).

The draft documents mention some of the ways that other LDPs can operate. For example, for explicit routing (with a source or controller setting up a Traffic Engineering path), it might be driven by two mechanisms known as CR-LDP (constraint-based routing) and RSVP-TE (extended RSVP driving an LDP).

A Concrete Example

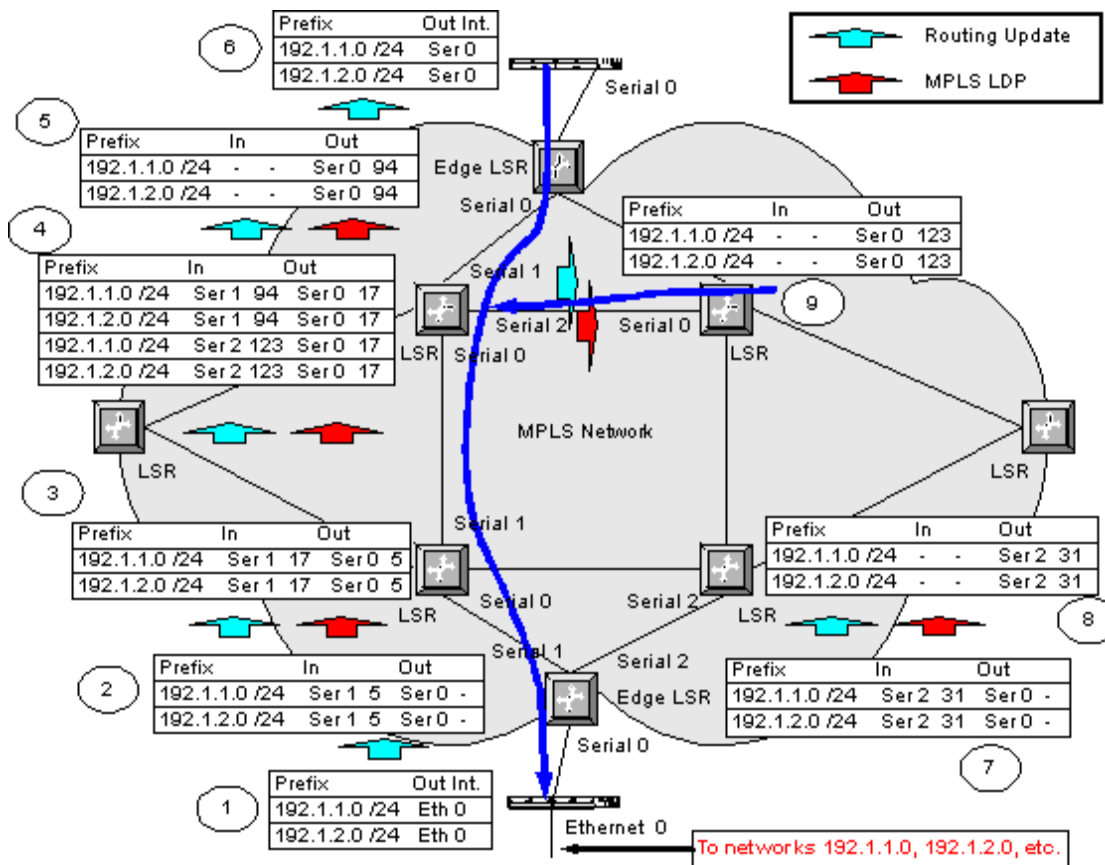
I don't know how you learn, but I always feel better when I see some of the details of how things work. For MPLS, the labels are the mysterious part. So let's look at how they are established and used in a concrete example. The discussion here will refer to the following diagram. I know it's a busy diagram. The numbers in circles refer to steps in the following explanation.

Step 1 (at the bottom). The bottom non-MPLS (customer) router has Class C networks 192.1.1.0 /24, 192.1.2.0 /24 somewhere out the Ethernet 0 interface. They are either directly connected (with a secondary address on the interface) or learned from another router. The table to the left of the bottom router attempts to suggest the routing table, which tracks the routing prefix, the outgoing interface, next hop router, and perhaps other information. The light blue arrow suggests that an ordinary routing update (you pick the protocol) advertises the routes to the Edge LSR above.

Step 2: The routes are advertised to the LSR above and to the left of the Edge LSR. Using LDP, the router selects a free (unused) label, 5, and advertises it to the upstream neighbor. The hyphen in the Out column is intended to note that all labels are to be popped (removed) in forwarding to the non-LSR below. Thus, a frame received on Serial 1 with label 5 is to be forwarded out Serial 0 with no label. The red arrow is intended to suggest LDP communicating the use of label 5 to the upstream LSR.

Step 3. The LSR has learned routes to the two prefixes we're tracking. It advertises the routes upstream. When LDP information is received, it records use of label 5 on outgoing interface Serial 0 for the two prefixes we're tracking. It then allocates label 17 on Serial 1 for this FEC, and uses LDP to communicate this to the upstream LSR. Thus, when label 17 is received on Serial 1, it is replaced with label 5 and the frame sent out Serial 0.

Steps 4 and 5: Proceed similarly. Note that there will be no labels received at the top Edge LSR, since the top router is not an MPLS participant, as we can see from its routing table (no labels!) in Step 6. The dark blue arrow shows the Label Switch Path (LSP) that has now been established. The table for Step 4 is bigger since this LSR has sent routing and LDP information to the LSR to its right.



Step 7: A routing advertisement might also be sent out interface Serial 2 from the Edge LSR at the bottom of the picture. It too can use LDP to tell the upstream LSR to use label 31 to deliver packets rapidly to the destinations we're tracking here.

Step 8: This LSR has perhaps not yet had time to propagate the routing information and label bindings upstream (or your author was getting fatigued).

Step 9: Here we have bindings that have passed from the left LSR to the right one. The right one uses label 123 for our two prefixes. Note that multiple flows can end up merging: frames bearing label 94 on Serial 1 or label 123 on Serial 2 all get relabelled with label 17 and sent out Serial 0. This indicates the multipoint-to-point behavior of IP MPLS.

Please note that the above example is incomplete, in that we have not yet propagated routes and bindings to all neighbors. You could see the same results after routing convergence if the metrics favored the links on the left side of the drawing. So don't attempt to read too much into this story. I'm just trying to show how routing and labels get propagated, and how the hop-by-hop behavior of LDP can still result in a Label Switch Path being established.

More Details and VC Merge

On ATM, the above behavior isn't quite what happens. The issue is that if you forward cells from two frames along the merged path shown in the figure above, you might intermingle the cells. Recall that data transmission uses ATM AAL5 encapsulation, and that there is no way to separate out intermingled cells from different frames in AAL5.

One solution is for the ATM switches and LSRs to do VC merge: know enough to delay the cells from one frame while cells from a different frame are transiting through the switch. This does create some interesting buffering and store-and-forward issues. Another approach is VP merge, where a common VPI but different VCI are used, to allow the edge LSR to sort the cells out. Yet another approach is to change the behavior of LDP over ATM, and have the upstream LSR drive the creation of the LSP. This results in separate VCs from every ingress LSR to the egress LSR, which may not scale well in terms of number of VCs.

See the references for more information on this.

Conclusion

For more to read about MPLS, there are the RFC and IETF draft documents. These are fairly readable, and can be found at the links at the beginning of this article.

There are two books available:

- MPLS: Technology and Applications by Bruce S. Davie, Yakov Rekhter (Morgan Kaufmann Press): <http://www.amazon.com/exec/obidos/ASIN/1558606564/>
- Multiprotocol Label Switching and Virtual Private Networks (Cisco Networking Fundamentals) by Ivan Pepelnjak, Jim Guichard (Cisco Press): <http://www.amazon.com/exec/obidos/ASIN/1587050021/>

The second is probably more focussed on MPLS as applied to VPNs, which is a fit topic for another article. (No outline online yet).

You might also wish to look at the MPLS Forum: <http://www.mplsforum.org> , and MPLS Resource Center: <http://www.mplsresource.com/> . This ought to get you thoroughly started.

Cisco white paper: http://www.cisco.com/warp/public/cc/cisco/mkt/wan/ipatm/tech/mpls_wp.htm

You might also look at the Cisco WAN switching documentation, which is actually a pretty good introduction to MPLS as well:

http://www.cisco.com/univercd/cc/td/doc/product/wanbu/9_3/mpls/index.htm . If you have the bandwidth, pull the PDF file and skim the first two hundred pages. **Cisco moved the link (as of 11/6/2001). The new link:**

http://www.cisco.com/univercd/cc/td/doc/product/wanbu/bpx8600/mpls/9_3_1/mpls.pdf

Some good Cisco documentation links to get you started:

- http://www.cisco.com/univercd/cc/td/doc/product/software/ios121/121cgcr/switch_c/xcprt4/index
- http://www.cisco.com/univercd/cc/td/doc/product/atm/c8540/wa5/12_0/3a_11/config/tag.htm
- http://www.cisco.com/univercd/cc/td/doc/product/software/ios121/121cgcr/switch_c/xcprt4/xcdba
- http://www.cisco.com/univercd/cc/td/doc/product/software/ios121/121cgcr/switch_r/index.htm

If you're wondering how to configure MPLS, well, turning on basic MPLS isn't that complicated. (Understanding it, well that's all clear now, isn't it?) We may look at basic configuration, and also at the fancier aspects of MPLS in one or more future articles. Fancier MPLS includes: QoS (CoS) with MPLS, also Traffic Engineering and VPNs.

Dr. Peter J. Welcher (CCIE #1773, CCSI #94014) is a Senior Consultant with Chesapeake NetCraftsmen. NetCraftsmen is a high-end consulting firm and Cisco Premier Partner dedicated to quality consulting and knowledge transfer. NetCraftsmen has nine CCIE's, with expertise including large network high-availability routing/switching and design, VoIP, QoS, MPLS, network management, security, IP multicast, and other areas. See <http://www.netcraftsmen.net> for more information about NetCraftsmen. Pete's links start at <http://www.netcraftsmen.net/welcher> . New articles will be posted under the Articles link. Questions, suggestions for articles, etc. can be sent to pjw@netcraftsmen.net .

8/7/2000

Copyright (C) 2000, Peter J. Welcher