INGREDIENTS FOR BETTER ROUTING? Read the Label

Christopher Metz, *IBM Corp.*

Administrators of wide-area IP networks—whether for Internet Service Providers or large corporate intranets—will face new and difficult challenges over the next few years. The growth of IP traffic is dwarfing the flat or modest growth rates of legacy (SNA, IPX) traffic. End-users are demanding consistently reliable performance for their network-based applications. Network managers are under pressure to enhance their infrastructures to support special or premium services for customers willing to pay for them. More bandwidth, lower network latencies, and high-speed traffic classification will be required to adequately support increased traffic volumes and new services.

IP Routing Latencies

Given the structure of current IP networks, the question is whether they will be able to meet these challenges. IP networks are built using routers, which in turn use protocols such as OSPF (open shortest path first) to exchange topology and reachability information so that packets are forwarded to the correct destination. Routers forward packets on a hop-by-hop basis by examining and processing information contained in each packet header.

The actions performed on the header at each router hop include

- determining the next hop address from a routing table lookup,
- decrementing the Time-To-Live field,
- computing a new header checksum, possibly fragmenting the packet,
- encapsulating the packet in the appropriate media on the outbound link, and
- sending the packet on its way.

The routing table lookup is the most time-consuming of these processes because it requires a "longest match" comparison between a variable-length network prefix that is part of the 32-bit IPv4 destination address and a corresponding entry in the router's routing table. Various lookup algorithms have been implemented in routers, but the number of memory accesses—and therefore the processing time—increases with the number of entries in the routing table.

Even if newer algorithms reduce or even bound the number of memory lookups, there is still a latency problem caused by the other actions that must be performed on each packet.[1] In addition, routers may soon be called upon to inspect even more packet information, such as source IP address, to determine special or nondefault handling.[2]

Today's Switched Backbone

To respond to the dramatic growth in traffic volumes, network providers have deployed connection-oriented NBMA (nonbroadcast multiaccess) switching technologies such as Frame Relay and ATM in the core of their networks. Switching provides high port densities and scalable bandwidth at a reasonable cost. It achieves the performance gain by using a small fixed-length connection identifier (label) in each packet/cell to index a local label-swap table in the switch hardware; the label-swap table contains a list of labels associated with the inbound port and the outbound ports, respectively. As the packet/cell travels through the switch fabric, the label associated on the inbound port is replaced by a label on the outbound port.

The label-swap tables are built when a virtual connection (VC) is established between two switched-network endpoints, such as routers. This notion of a VC in which hardware switch resources (labels, buffers, and so on) are allocated on every switch between the two VC endpoints, enables Frame Relay and ATM to provide reliable performance and even QoS guarantees. Moreover, network providers can place specific customer traffic flows on a separate VC and then provision it to pass through a specific set of links and switches.

Integrating Routing and Switching

IP routing has proved necessary to building and supporting scalable networks, and switching provides the high-performance machinery and per-customer connections needed to support large traffic volumes and new services. Integrating the two has been a goal of network engineering over the past five years.

IP Over NBMA. One way to accomplish this is to interconnect all routers with VCs. Figure 1a illustrates this approach, called IP over NBMA. To achieve the full benefits of end-to-end switch performance, all routers must be connected in a full mesh, thus requiring $O(N^2)$ virtual connections where N is the number of routers. As N grows, the number of VCs grows exponentially, which can impose an overhead burden on several fronts.

First, the number of VCs needed to maintain a full mesh could approach the supported VC threshold of the switching fabric. (This is more an

implementation than architectural statement, because switches on the market today support a finite number of connections.)

Second, as you might guess, routing protocols do not scale well when operating over an NBMA network. This is because a VC is treated as a physical link that forms an adjacency with another router. The overhead of processing routing-protocol updates and computing routes is proportional to the number of links in the network and router adjacencies. So while the IP over NBMA approach is in use in many networks today, there are several legitimate scaling concerns that must be addressed.

NHRP Shortcut Path. Another solution does away with the full mesh requirement and instead exploits the dynamic switched-virtual-connection (SVC) capability of Frame Relay and, in particular, ATM. The idea here is to support hop-by-hop routing but also allow a router to query for a next-hop address closer to the destination. A "shortcut" path that bypasses intermediate router hops can then be dynamically established.

Building the shortcut path is accomplished by an ingress router first detecting a particular traffic flow that can benefit from a shortcut path, and then forwarding a query message toward the destination. The last-hop or egress router on the ATM network should respond to the query and return its own ATM address to the ingress router. The ingress router can then establish an SVC to the egress router that forms the shortcut path, as shown in Figure 1b.

The IP Over NBMA (ION) working group of the IETF, which has addressed this problem for the past few years, recently completed work on the Next-Hop Resolution Protocol. The ATM Forum has incorporated **NHRP** and related elements in its Multiprotocol over ATM (MPOA) specification, which supports a network-layer routing function on top of an ATM network.[3] While presumably consuming fewer VC resources than the IP over NMBA approach due to the use of SVCs, NHRP/MPOA has its own operational and scaling issues. First, it depends on an additional query-based control protocol for resolving IP and NBMA addresses. All routers along the routed path must be able to interpret NHRP messages and maintain a cache of IP/ATM address bindings. The shortcut paths are unicast only, and of course there is some latency in establishing them. And finally, persistent routing loops can form under certain topological conditions.

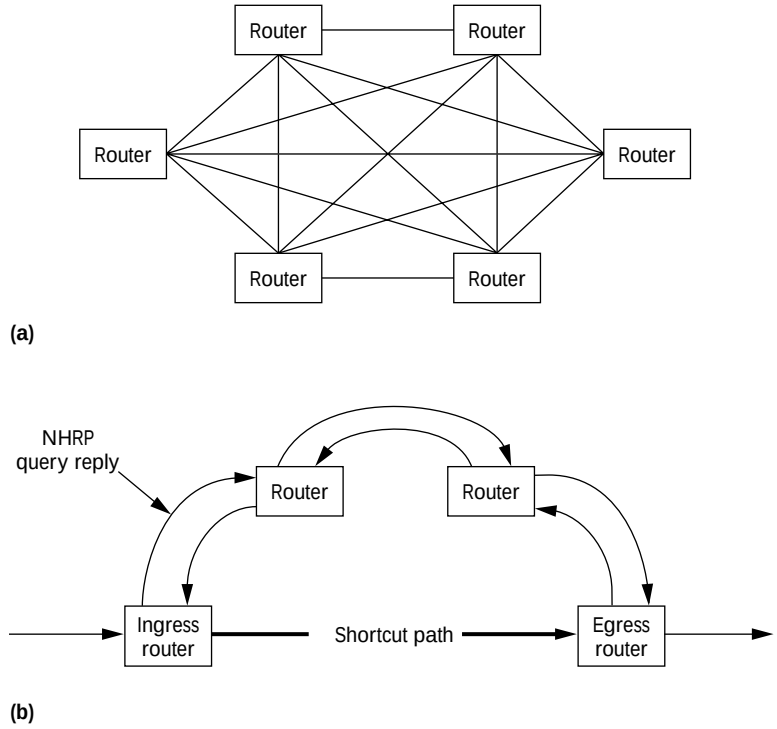

Figure 1. Nonbroadcast multiaccess (NBMA) switching technologies: (a) IP Over NBMA integrates the scalability of IP routing with the high performance associated with end-to-end switching by connecting each router in a full mesh. (b) Next-Hop Resolution Protocol from the IETF NBMA working group implements the shortcut path approach, bypassing intermediate hops.

IP Switching

Both the IP over NBMA and MPOA/NHRP solutions retain the functions and services provided by the underlying Layer 2 switching fabric (usually ATM), while allowing the IP routing topology and protocols to be logically placed on top. This so-called *overlay model* presents some additional complexity in that two topologies (router and switches), two address spaces (IP and ATM), and two routing protocols (for example, OSPF for IP and PNNI for ATM) must be supported and maintained.

A simplified approach for the integration of routing and switching is to remove the Layer 2 switching topology, protocols, and address spaces in favor of IP, but to retain the packet/cell switching hardware. This outfits the IP forwarding device with a switch fabric that can forward packets/cells at the same high-performance rates as traditional switches.[4]

In fact, this is the same label-swapping forwarding technique used to switch Frame Relay PDUs and ATM cells. So whereas the network topology consisted of separate routing and switching

devices in the overlay model, it now appears as a single topology of hybrid switch-router devices that run standard IP routing protocols and use switching to forward packets/cells through the network. However, the *peer model*, as this is called, does require an additional control component operating on the IP switch to coordinate which packets will be redirected through switched hardware and which packets will be processed at Layer 3.

Using switching hardware, particularly ATM, to improve IP routing performance spawned a whole new suite of integrated routing-switching solutions collectively known as *IP switching*.[5]

Broadly speaking, IP switching can be defined as the protocols and machinery that use Layer 2 switching to accelerate the forwarding of IP packets.

IP switching solutions can be generally classified as flow- or topology-driven.

Flow-Driven. Flow-driven solutions select and then redirect individual IP flows through switching hardware, where a flow is defined as a sequence of packets that share a common source/destination IP address and port number. Each IP switch along the routed path first classifies a particular flow by examining multiple fields in each packet and then instructs its neighboring IP switch to re-label the cells of the flow with a new VPI/VCI (virtual path identifier/virtual connection identifier) value.

Once the relabeling process occurs for a particular flow on the incoming and outgoing ports of an IP switch, it is simple to splice or concatenate the incoming and outgoing entries in the label-swap table, thus forming an internal cell-switched path. Each IP switch along the routed path will perform this function on a per-flow basis, thus forming an ingress-to-egress cell switched path. What started out initially as a hop-by-hop routed traffic flow ends up as an ATM-switched traffic flow.

A start-up formerly known as **Ipsilon** (since purchased by Nokia) first popularized the concept of IP switching and in fact published their specific IP switching protocols as informational RFCs. Toshiba has also developed a flow-driven IP switching solution and platform that is unique in its ability to operate in conjunction with stand-alone ATM switches.

What started out initially as a hop-by-hop routed traffic flow ends up as an ATM-switched traffic flow.

Topology-Driven. Topology-driven IP switching is based on the IP network topology maintained by routing protocols running in the IP switch. New VPI/VCI labels associated with a specific destination IP prefix (called route/label bindings) are generated and distributed to the other IP switches in the routing domain. All traffic destined for a particular network will then follow the switched path based on the new VPI/VCI values.

Solutions that work on this basis were first introduced by Cisco with **Tag Switching** and then by IBM with Aggregate Route-Based IP Switching (Aris). These solutions also introduced some important innovations to the whole class of IP switching solutions.

Tag Switching extended the IP switching model to work over different data-link technologies while retaining the performance advantages and functional independence of label-swapping and forwarding. When a specific data-link encapsulation does not include an inherent swappable label in the header (such as PPP or Ethernet), a tag shim header is inserted in between the native Layer 2 and IP header. A tag switching router (TSR) can then forward the packet on the basis of the tag contents—not the variable length destination prefix in the IP header.[6]

Aris introduced the concept of VC merging that allows for multiple upstream VCs to be merged into a single downstream VC without interleaving cells from different frames.[7] This allows the creation of multipoint-to-point switched paths. VC merging improves scalability. Rather than requiring $O(N^2)$ connections to support a switched path between N number of routers, a network capable of switched path merging only requires $O(N)$ multipoint-to-point connections to support a switched path between all routers.

Multiprotocol Label Switching

Given the plethora of similar but vendor-proprietary IP switching schemes that appeared by the end of 1996, it did not take long for many in the industry to conclude that a single standard for integrated routing and switching would be advantageous. The IETF formed a working group and christened the effort Multiprotocol Label Switching.[8]

The name is derived from the use of simple label swapping as the forwarding mechanism of choice. The primary objective of the **MPLS** working group is to develop a solution for integrated routing and switching that lowers the price and performance of routing, improves the scalability of IP routing vis-à-vis the IP over NBMA model, and facilitates the delivery of new networking services while retaining the use of a simple, scalable forwarding mechanism.

Label Switch Routers. The MPLS term for an IP switch is a Label Switch Router. Like a generic IP switch, an LSR can forward packets at Layer 3 and switch packets at Layer 2. It operates traditional IP

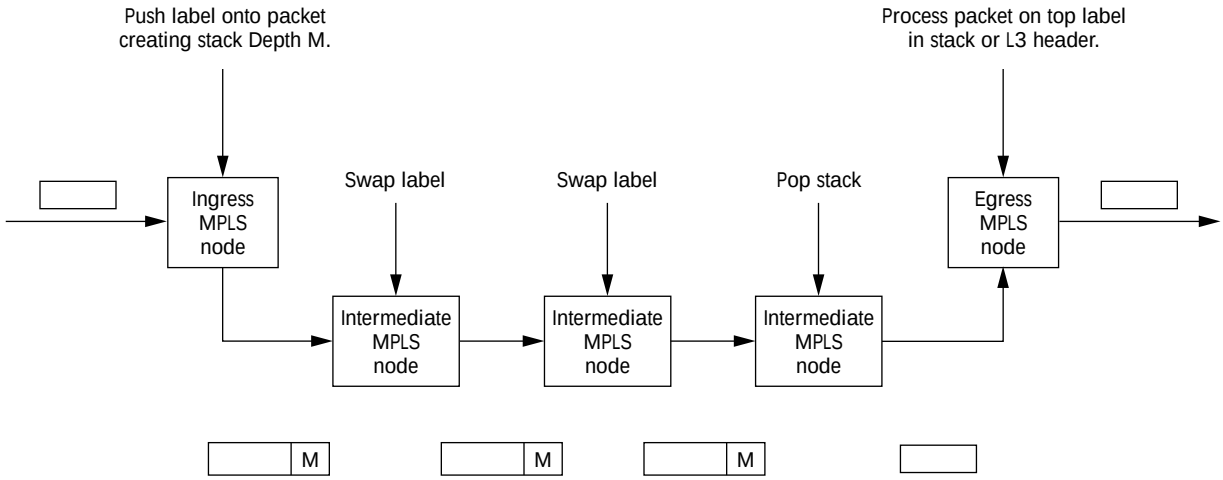

Figure 2. Forwarding a packet along a label-switched path. At the LSP ingress, a label is pushed into the packet, creating a stack of labels with a depth M. Intermediate LSRs receive and process the packet. Only the top label in the stack is acted upon by swapping the label with a new label corresponding to the next-hop downstream LSR. At the LSP egress, the LSR has to "pop" the stack to get to the next label.

routing protocols and may run a specialized control protocol to coordinate FEC (forwarding equivalence class)/label bindings with neighboring LSRs. An FEC is a group of IP packets that are forwarded over the same path and treated in the same manner, and can therefore be mapped to a single label by an LSR.

To perform label swapping, an LSR must also maintain a Label Information Base. A LIB is a connection table consisting of inbound and outbound ports and associated labels.

MPLS Labels. A label is a short, fixed-length value that is contained in each packet and is used to forward the packet through the network. A pair of LSRs must agree up front on the value and meaning of the label.

For example, the downstream LSR (in the context of the traffic flow) will tell the upstream LSR that a particular label *X* will be used to represent a FEC called *A*." A label has significance, therefore, only between a pair of communicating LSRs, and represents the packets belonging to a particular FEC flowing from the upstream to the downstream LSR.

MPLS can support labels that are appended to existing frame or packet structures, such as Ethernet or PPP, or it can use label structures contained in the data-link layer (Frame Relay, ATM).

Label-Swapping/Forwarding. MPLS includes a mechanism to support label-swapping/forwarding—a simple, fast procedure that is also used in Frame Relay and ATM switches.

Unlike conventional IP routing, label-swapping forwarding does not require analysis of variable-length portions of the header's contents. Once the labels for a particular FEC have been distributed among the LSRs along the packet route, a label-switched path (LSP) is built from the network ingress to the egress (see Figure 2).

When packets enter the network, the ingress LSR examines multiple fields in the packet header to determine what FEC the packet belongs to. If an FEC/ label binding is present, the ingress LSR affixes a label to the packet and directs it to the appropriate outbound interface. The packet (or cell, in the case of ATM data links) is then label-swapped through the network until it reaches the egress LSR, where the label is removed and the packet is processed based on its header contents.

This approach improves performance by pushing per-packet IP header analysis and processing to the edge (ingress) and performing them only once rather than at each intermediate hop. Intermediate processing consists of matching a short, fixed-length label in the packet with a corresponding entry in the LSR's LIB and then swapping the labels—a ranking process typically performed in high-performance hardware-based switches. A packet may contain more than one label (as it would, for example, when a previously labeled packet is tunneled through an LSP); however, the LSR is only required to process the top label in the stack.

The LSR at the egress of the LSP will make a forwarding decision based on the contents of the next lower label in the stack. This means that the egress LSR must "pop" the stack to get to the next label.

A subtle optimization to this process, called *penultimate hop popping*, can be achieved if the egress node tells the next to last LSR in the LSP to pop the stack (accomplished through the use of an Implicit Null label). The packet then arrives at the egress LSR with the label that will be used to for-

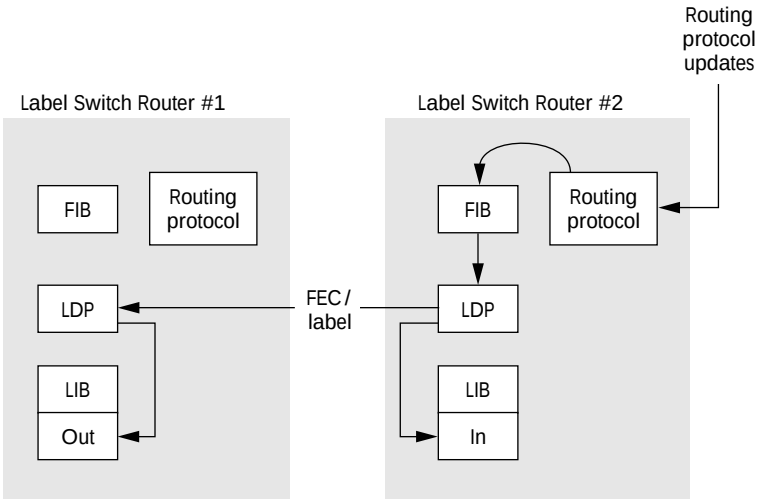

Figure 3. Topology-driven label assignment. LSR #2 receives routing-table updates that trigger new path calculations resulting in new routing-table entries (FIB). For each FEC, LSR #2 allocates a label on the inbound port and places it in the Label Information Base. The label and associated FEC are communicated upstream to LSR #1, which places the label in a corresponding outbound port entry in its LIB.

ward the packet already at the top, saving the egress device from having to do an extra table lookup.

Label Distribution. Another core MPLS technology is label distribution, the process of distributing FEC/label bindings among participating LSRs. This can be accomplished by a separate label distribution protocol (LDP) or by piggy-backing the FEC/label bindings in existing control protocols such as RSVP and BGP).

The basic operation is for the downstream LSR to allocate the label and then distribute the FEC/label binding to the adjacent upstream LSR. In the situation where the FEC corresponds to an address prefix distributed by a dynamic routing protocol, the formation of an LSP (using either LDP or another control protocol) can be done in an independent or ordered manner.

In the case of *independent LSP control*, an LSR makes an independent decision to bind a label to an FEC and communicates this binding to its LSR neighbor(s). In *ordered LSP control*, the LSR will only bind a label to an FEC if it is the LSP egress or if it has already received a binding for the specified FEC from the next-hop LSR. Ordered LSP control is used to build a path with nondefault properties. Paths that pass through a specified sequence of nodes or paths that must be loop-free are two examples of paths with non-default properties. The MPLS architecture is capable of supporting both independent and ordered LSP control.

MPLS LSPs can be built based on the arrival of specific data flows (flow-driven), reservation setup messages (such as RSVP), or routing table update messages (topology-driven). Given the major requirement of scalability and the MPLS design focus on very large IP networks, the topology-driven approach will be deployed most often. Figure 3 illustrates conceptually how the presence of routing table updates initiates the exchange of FEC-label bindings between a pair of LSRs.

Network Service Over MPLS

In MPLS, it is possible for the ingress router to map the packet(s) to any number of different FECs. For example, an FEC may be based on the destination network address, a group of destination addresses, a source/destination address pair, a source address only or even the physical point of entry into the network. A FEC may also represent all packets that are to traverse an explicit nondefault path. Independent of whatever complex policy is invoked to assign a packet to a FEC, the forwarding of the packet through the network is still based on label swapping. Thus, MPLS facilitates the use of policy-based routing in a much simpler and more straightforward manner then would otherwise be possible using traditional IP forwarding.

Given its ability to associate any type of FEC with a label and LSP, MPLS enables a number of different and interesting network services to be deployed. One of the most useful services provided by MPLS is the ability to perform traffic engineering. Traffic engineering, loosely defined, is the ability to direct traffic flows over separate, nondefault, explicitly defined paths. Instead of depending on a dynamic IP routing protocol to compute a path based on a single metric (number of hops or lowest link cost) for all packets sharing a common destination or egress point, MPLS can be used to build separate LSPs for separate packet flows that may share a common ingress and egress node. Directing traffic over nondefault explicit paths incurs no performance penalty because the forwarding mechanism for default and nondefault LSPs is the same.

Traffic engineering is a useful tool for network providers for several reasons. Aggregate network traffic loads can be balanced across all available backbone links to respond to changing network usage patterns or to avoid network bottlenecks. It can also be used to offer a value-add service with more bandwidth and lower delays by directing a particular customer's traffic over a separate path. LDP and RSVP have both been proposed as the setup protocol for building explicit route LSPs (ER-LSP).

MPLS and ATM

When running on ATM hardware, both MPLS and ATM Forum protocols use the same packet formats (53-byte cells), same labels (VPI/VCI), and same label-swapping cell-switching forwarding mechanisms. The fundamental differences lie in the fact that MPLS has no use for ATM addressing, ATM routing, or ATM Forum protocols, but instead uses IP addressing, dynamic IP routing, and an additional control protocol (LDP) to map FECs to labels and thus form LSPs.

In general, MPLS is concerned solely with creating and distributing FEC/label mappings so that IP traffic can be forwarded more efficiently through a network over a default or nondefault path. When operating on ATM hardware, MPLS is just another ATM control plane, albeit one driven primarily by IP routing protocols and designed exclusively for the benefit of routed IP traffic.

In practice, MPLS will most likely coexist with native ATM in a "ships in the night" configuration, where the two operate in either a mutually exclusive or an integrated manner (when MPLS nodes communicate through native ATM switches).

Finally, it should be pointed out that the MPLS architecture can operate over any data-link technology—not just ATM. A network provider can therefore configure and run MPLS in a single domain consisting of PPP, Frame Relay, ATM, and broadcast LAN data-links.

Conclusion

There are still some issues to be resolved within the IETF working group before MPLS becomes a full standard, but established industry veterans like **Cisco** and **Ericsson**, as well as startups such as **Juniper** and **Ennovate**, have announced support for MPLS. ■

REFERENCES

1. M. Waldvogel et al., "Scalable High-Speed IP Routing Lookups," *ACM/Sigcomm Proc.*, Fall 1997.
2. See, for example, IETF Differentiated Services Working Group charter, http://www.ietf.org/html.charters/diffserv-charter.html.
3. "Multi-Protocol over ATM Specification V1.0," ATM Forum, af-mpoa-0087.000, July 1997; available at ftp://ftp.atmforum.com/ pub/approved-specs/af-mpoa-0087.000.pdf.
4. P.P. White, "ATM Switching and IP Routing Integration: The Next Stage in the Internet Evolution," *IEEE Comm.*, Vol. 36, No. 4, Apr.l 1998.
5. C. Metz, *IP Switching: Protocols and Architectures*," McGraw-Hill, 1998.
6. Y. Rekhter et al., "Tag Switching Architecture Overview," *IEEE Proc.*, Vol. 85, No. 12, Dec. 1997.
7. N. Feldman et al., "Aggregate Route-Based IP Switching (ARIS)," IBM Technical Report 29.2353.
8. A. Viswanathan et al., "Evolution of Multiprotocol Label Switching," *IEEE Comm.*, Vol. 36, No. 5, May 1998, pp. 165-173.

Chris Metz is an IP technology consultant for IBM, based in Raleigh, N.C. He specializes in advanced and emerging IP technologies and protocols. He is coauthor of *ATM and Multiprotocol Networking* (McGraw-Hill, 1997) and author of *IP Switching: Protocols and Architectures* (McGraw-Hill, 1998). Metz is a member of IEEE and ACM/SIGComm. Readers may contact Metz at metzcy@ibm.net

URLs for On the Wire

Ennovate • www.ennovatenetworks.com/
Ericsson • www.ericsson.se/
Ipsilon • www.ipsilon.com/about/technology/
Juniper • www.junipernetworks.com/
MPLS • www.employees.org/~mpls/
NHRP • www.ietf.org/html.charters/ion-charter.html
Tag Switching • www.cisco.com/warp/public/732/tag/index.html