
Network Working Group
Request for Comments: 2105
Category: Informational

Y. Rekhter
B. Davie
D. Katz
E. Rosen
G. Swallow
Cisco Systems, Inc.

February 1997

Cisco Systems' Tag Switching Architecture Overview

Status of this Memo

This memo provides information for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

IESG Note:

This protocol is NOT the product of an IETF working group nor is it a standards track document. It has not necessarily benefited from the widespread and in depth community review that standards track documents receive.

Abstract

This document provides an overview of a novel approach to network layer packet forwarding, called tag switching. The two main components of the tag switching architecture - forwarding and control - are described. Forwarding is accomplished using simple label-swapping techniques, while the existing network layer routing protocols plus mechanisms for binding and distributing tags are used for control. Tag switching can retain the scaling properties of IP, and can help improve the scalability of IP networks. While tag switching does not rely on ATM, it can straightforwardly be applied to ATM switches. A range of tag switching applications and deployment scenarios are described.

Table of Contents

1. Introduction

Continuous growth of the Internet demands higher bandwidth within the Internet Service Providers (ISPs). However, growth of the Internet is not the only driving factor for higher bandwidth - demand for higher bandwidth also comes from emerging multimedia applications. Demand for higher bandwidth, in turn, requires higher forwarding performance (packets per second) by routers, for both multicast and unicast traffic.

The growth of the Internet also demands improved scaling properties of the Internet routing system. The ability to contain the volume of routing information maintained by individual routers and the ability to build a hierarchy of routing knowledge are essential to support a high quality, scalable routing system.

We see the need to improve forwarding performance while at the same time adding routing functionality to support multicast, allowing more flexible control over how traffic is routed, and providing the ability to build a hierarchy of routing knowledge. Moreover, it becomes more and more crucial to have a routing system that can support graceful evolution to accommodate new and emerging requirements.

Tag switching is a technology that provides an efficient solution to these challenges. Tag switching blends the flexibility and rich functionality provided by Network Layer routing with the simplicity provided by the label swapping forwarding paradigm. The simplicity of the tag switching forwarding paradigm (label swapping) enables improved forwarding performance, while maintaining competitive price/performance. By associating a wide range of forwarding granularities with a tag, the same forwarding paradigm can be used to support a wide variety of routing functions, such as destination-based routing, multicast, hierarchy of routing knowledge, and flexible routing control. Finally, a combination of simple forwarding, a wide range of forwarding granularities, and the ability to evolve routing functionality while preserving the same forwarding paradigm enables a routing system that can gracefully evolve to

accommodate new and emerging requirements.

The rest of the document is organized as follows. Section 2 introduces the main components of tag switching, forwarding and control. Section 3 describes the forwarding component. Section 4 describes the control component. Section 5 describes how tag switching could be used with ATM. Section 6 describes the use of tag switching to help provide a range of qualities of service. Section 7 briefly describes possible deployment scenarios. Section 8 summarizes the results.

2. Tag Switching components

Tag switching consists of two components: forwarding and control. The forwarding component uses the tag information (tags) carried by packets and the tag forwarding information maintained by a tag switch to perform packet forwarding. The control component is responsible for maintaining correct tag forwarding information among a group of interconnected tag switches.

3. Forwarding component

The fundamental forwarding paradigm employed by tag switching is based on the notion of label swapping. When a packet with a tag is received by a tag switch, the switch uses the tag as an index in its Tag Information Base (TIB). Each entry in the TIB consists of an incoming tag, and one or more sub-entries of the form (outgoing tag, outgoing interface, outgoing link level information). If the switch finds an entry with the incoming tag equal to the tag carried in the packet, then for each (outgoing tag, outgoing interface, outgoing link level information) in the entry the switch replaces the tag in the packet with the outgoing tag, replaces the link level information (e.g MAC address) in the packet with the outgoing link level information, and forwards the packet over the outgoing interface.

From the above description of the forwarding component we can make several observations. First, the forwarding decision is based on the exact match algorithm using a fixed length, fairly short tag as an index. This enables a simplified forwarding procedure, relative to longest match forwarding traditionally used at the network layer. This in turn enables higher forwarding performance (higher packets per second). The forwarding procedure is simple enough to allow a straightforward hardware implementation.

A second observation is that the forwarding decision is independent of the tag's forwarding granularity. For example, the same forwarding algorithm applies to both unicast and multicast - a unicast entry would just have a single (outgoing tag, outgoing interface, outgoing link level information) sub-entry, while a multicast entry may have one or more (outgoing tag, outgoing interface, outgoing link level information) sub-entries. (For multi-access links, the outgoing link level information in this case would include a multicast MAC address.) This illustrates how with tag switching the same forwarding paradigm can be used to support different routing functions (e.g., unicast, multicast, etc...)

The simple forwarding procedure is thus essentially decoupled from the control component of tag switching. New routing (control) functions can readily be deployed without disturbing the forwarding paradigm. This means that it is not necessary to re-optimize forwarding performance (by modifying either hardware or software) as new routing functionality is added.

3.1. Tag encapsulation

Tag information can be carried in a packet in a variety of ways:

- as a small "shim" tag header inserted between the layer 2 and the Network Layer headers;

- as part of the layer 2 header, if the layer 2 header provides adequate semantics (e.g., ATM, as discussed below);

- as part of the Network Layer header (e.g., using the Flow Label field in IPv6 with appropriately modified semantics).

It is therefore possible to implement tag switching over virtually any media type including point-to-point links, multi-access links, and ATM.

Observe also that the tag forwarding component is Network Layer independent. Use of control component(s) specific to a particular Network Layer protocol enables the use of tag switching with different Network Layer protocols.

4. Control component

Essential to tag switching is the notion of binding between a tag and Network Layer routing (routes). To provide good scaling characteristics, while also accommodating diverse routing functionality, tag switching supports a wide range of forwarding granularities. At one extreme a tag could be associated (bound) to a group of routes (more specifically to the Network Layer Reachability Information of the routes in the group). At the other extreme a tag could be bound to an individual application flow (e.g., an RSVP flow). A tag could also be bound to a multicast tree.

The control component is responsible for creating tag bindings, and then distributing the tag binding information among tag switches. The control component is organized as a collection of modules, each designed to support a particular routing function. To support new routing functions, new modules can be added. The following describes some of the modules.

4.1. Destination-based routing

In this section we describe how tag switching can support destination-based routing. Recall that with destination-based routing a router makes a forwarding decision based on the destination address carried in a packet and the information stored in the Forwarding Information Base (FIB) maintained by the router. A router constructs its FIB by using the information the router receives from routing protocols (e.g., OSPF, BGP).

To support destination-based routing with tag switching, a tag switch, just like a router, participates in routing protocols (e.g., OSPF, BGP), and constructs its FIB using the information it receives from these protocols.

There are three permitted methods for tag allocation and Tag Information Base (TIB) management: (a) downstream tag allocation, (b) downstream tag allocation on demand, and (c) upstream tag allocation. In all cases, a switch allocates tags and binds them to address prefixes in its FIB. In downstream allocation, the tag that is carried in a packet is generated and bound to a prefix by the switch at the downstream end of the link (with respect to the direction of data flow). In upstream allocation, tags are allocated and bound at the upstream end of the link. 'On demand' allocation means that tags will only be allocated and distributed by the downstream switch when it is requested to do so by the upstream switch. Methods (b) and (c) are most useful in ATM networks (see Section 5). Note that in downstream allocation, a switch is responsible for creating tag bindings that apply to incoming data packets, and receives tag bindings for outgoing packets from its neighbors. In upstream allocation, a switch is responsible for creating tag bindings for outgoing tags, i.e. tags that are applied to data packets leaving the switch, and receives bindings for incoming tags from its neighbors.

The downstream tag allocation scheme operates as follows: for each route in its FIB the switch allocates a tag, creates an entry in its Tag Information Base (TIB) with the incoming tag set to the allocated tag, and then advertises the binding between the (incoming) tag and the route to other adjacent tag switches. The advertisement could be accomplished by either piggybacking the binding on top of the existing routing protocols, or by using a separate Tag Distribution

Protocol [TDP]. When a tag switch receives tag binding information for a route, and that information was originated by the next hop for that route, the switch places the tag (carried as part of the binding information) into the outgoing tag of the TIB entry associated with the route. This creates the binding between the outgoing tag and the route.

With the downstream tag allocation on demand scheme, operation is as follows. For each route in its FIB, the switch identifies the next hop for that route. It then issues a request (via TDP) to the next hop for a tag binding for that route. When the next hop receives the request, it allocates a tag, creates an entry in its TIB with the incoming tag set to the allocated tag, and then returns the binding between the (incoming) tag and the route to the switch that sent the original request. When the switch receives the binding information, the switch creates an entry in its TIB, and sets the outgoing tag in the entry to the value received from the next hop.

The upstream tag allocation scheme is used as follows. If a tag switch has one or more point-to-point interfaces, then for each route in its FIB whose next hop is reachable via one of these interfaces, the switch allocates a tag, creates an entry in its TIB with the outgoing tag set to the allocated tag, and then advertises to the next hop (via TDP) the binding between the (outgoing) tag and the route. When a tag switch that is the next hop receives the tag binding information, the switch places the tag (carried as part of the binding information) into the incoming tag of the TIB entry associated with the route.

Once a TIB entry is populated with both incoming and outgoing tags, the tag switch can forward packets for routes bound to the tags by using the tag switching forwarding algorithm (as described in Section 3).

When a tag switch creates a binding between an outgoing tag and a route, the switch, in addition to populating its TIB, also updates its FIB with the binding information. This enables the switch to add tags to previously untagged packets.

To understand the scaling properties of tag switching in conjunction with destination-based routing, observe that the total number of tags that a tag switch has to maintain can not be greater than the number of routes in the switch's FIB. Moreover, in some cases a single tag could be associated with a group of routes, rather than with a single route. Thus, much less state is required than would be the case if tags were allocated to individual flows.

In general, a tag switch will try to populate its TIB with incoming and outgoing tags for all routes to which it has reachability, so that all packets can be forwarded by simple label swapping. Tag allocation is thus driven by topology (routing), not traffic - it is the existence of a FIB entry that causes tag allocations, not the arrival of data packets.

Use of tags associated with routes, rather than flows, also means that there is no need to perform flow classification procedures for all the flows to determine whether to assign a tag to a flow. That, in turn, simplifies the overall scheme, and makes it more robust and stable in the presence of changing traffic patterns.

Note that when tag switching is used to support destination-based routing, tag switching does not completely eliminate the need to perform normal Network Layer forwarding. First of all, to add a tag to a previously untagged packet requires normal Network Layer forwarding. This function could be performed by the first hop router, or by the first router on the path that is able to participate in tag switching. In addition, whenever a tag switch aggregates a set of routes (e.g., by using the technique of hierarchical routing), into a single tag, and the routes do not share a common next hop, the switch needs to perform Network Layer forwarding for packets carrying that tag. However, one could observe that the number of places where routes get aggregated is smaller than the total number of places where forwarding decisions have to be made. Moreover, quite often aggregation is applied to only a subset of the routes maintained by a tag switch. As a result, on average a packet can be forwarded most of the time using the tag switching algorithm.

4.2. Hierarchy of routing knowledge

The IP routing architecture models a network as a collection of routing domains. Within a domain, routing is provided via interior routing (e.g., OSPF), while routing across domains is provided via exterior routing (e.g., BGP). However, all routers within domains that carry transit traffic (e.g., domains formed by Internet Service Providers) have to maintain information provided by not just interior routing, but exterior routing as well. That creates certain problems. First of all, the amount of this information is not insignificant. Thus it places additional demand on the resources required by the routers. Moreover, increase in the volume of routing information quite often increases routing convergence time. This, in turn, degrades the overall performance of the system.

Tag switching allows the decoupling of interior and exterior routing, so that only tag switches at the border of a domain would be required to maintain routing information provided by exterior routing, while all other switches within the domain would just maintain routing information provided by the domain's interior routing (which is usually significantly smaller than the exterior routing information). This, in turn, reduces the routing load on non-border switches, and shortens routing convergence time.

To support this functionality, tag switching allows a packet to carry not one but a set of tags, organized as a stack. A tag switch could either swap the tag at the top of the stack, or pop the stack, or swap the tag and push one or more tags into the stack.

When a packet is forwarded between two (border) tag switches in different domains, the tag stack in the packet contains just one tag. However, when a packet is forwarded within a domain, the tag stack in the packet contains not one, but two tags (the second tag is pushed by the domain's ingress border tag switch). The tag at the top of the stack provides packet forwarding to an appropriate egress border tag switch, while the next tag in the stack provides correct packet forwarding at the egress switch. The stack is popped by either the egress switch or by the penultimate (with respect to the egress switch) switch.

The control component used in this scenario is fairly similar to the one used with destination-based routing. In fact, the only essential difference is that in this scenario the tag binding information is distributed both among physically adjacent tag switches, and among border tag switches within a single domain. One could also observe that the latter (distribution among border switches) could be trivially accommodated by very minor extensions to BGP (via a separate Tag Binding BGP attribute).

4.3. Multicast

Essential to multicast routing is the notion of spanning trees. Multicast routing procedures (e.g., PIM) are responsible for constructing such trees (with receivers as leafs), while multicast forwarding is responsible for forwarding multicast packets along such trees.

To support a multicast forwarding function with tag switching, each tag switch associates a tag with a multicast tree as follows. When a tag switch creates a multicast forwarding entry (either for a shared or for a source-specific tree), and the list of outgoing interfaces for the entry, the switch also creates local tags (one per outgoing interface). The switch creates an entry in its TIB and populates (outgoing tag, outgoing interface, outgoing MAC header) with this information for each outgoing interface, placing a locally generated tag in the outgoing tag field. This creates a binding between a multicast tree and the tags. The switch then advertises over each outgoing interface associated with the entry the binding between the tag (associated with this interface) and the tree.

When a tag switch receives a binding between a multicast tree and a tag from another tag switch, if the other switch is the upstream neighbor (with respect to the multicast tree), the local switch places the tag carried in the binding into the incoming tag component of the TIB entry associated with the tree.

When a set of tag switches are interconnected via a multiple-access subnetwork, the tag allocation procedure for multicast has to be coordinated among the switches. In all other cases tag allocation procedure for multicast could be the same as for tags used with destination-based routing.

4.4. Flexible routing (explicit routes)

One of the fundamental properties of destination-based routing is that the only information from a packet that is used to forward the packet is the destination address. While this property enables highly scalable routing, it also limits the ability to influence the actual paths taken by packets. This, in turn, limits the ability to evenly distribute traffic among multiple links, taking the load off highly utilized links, and shifting it towards less utilized links. For Internet Service Providers (ISPs) who support different classes of service, destination-based routing also limits their ability to segregate different classes with respect to the links used by these classes. Some of the ISPs today use Frame Relay or ATM to overcome the limitations imposed by destination-based routing. Tag switching, because of the flexible granularity of tags, is able to overcome these limitations without using either Frame Relay or ATM.

To provide forwarding along the paths that are different from the paths determined by the destination-based routing, the control component of tag switching allows installation of tag bindings in tag switches that do not correspond to the destination-based routing paths.

5. Tag switching with ATM

Since the tag switching forwarding paradigm is based on label swapping, and since ATM forwarding is also based on label swapping, tag switching technology can readily be applied to ATM switches by implementing the control component of tag switching.

The tag information needed for tag switching can be carried in the VCI field. If two levels of tagging are needed, then the VPI field could be used as well, although the size of the VPI field limits the size of networks in which this would be practical. However, for most applications of one level of tagging the VCI field is adequate.

To obtain the necessary control information, the switch should be able (at a minimum) to participate as a peer in Network Layer routing protocols (e.g., OSPF, BGP). Moreover, if the switch has to perform routing information aggregation, then to support destination-based unicast routing the switch should be able to perform Network Layer forwarding for some fraction of the traffic as well.

Supporting the destination-based routing function with tag switching on an ATM switch may require the switch to maintain not one, but several tags associated with a route (or a group of routes with the same next hop). This is necessary to avoid the interleaving of packets which arrive from different upstream tag switches, but are sent concurrently to the same next hop. Either the downstream tag allocation on demand or the upstream tag allocation scheme could be used for the tag allocation and TIB maintenance procedures with ATM switches.

Therefore, an ATM switch can support tag switching, but at the minimum it needs to implement Network Layer routing protocols, and the tag switching control component on the switch. It may also need to support some network layer forwarding.

Implementing tag switching on an ATM switch would simplify integration of ATM switches and routers - an ATM switch capable of tag switching would appear as a router to an adjacent router. That could provide a viable, more scalable alternative to the overlay model. It also removes the necessity for ATM addressing, routing and signalling schemes. Because the destination-based forwarding approach described in section 4.1 is topology driven rather than traffic driven, application of this approach to ATM switches does not high call setup rates, nor does it depend on the longevity of flows.

Implementing tag switching on an ATM switch does not preclude the ability to support a traditional ATM control plane (e.g., PNNI) on the same switch. The two components, tag switching and the ATM control plane, would operate in a Ships In the Night mode (with VPI/VCI space and other resources partitioned so that the components do not interact).

6. Quality of service

Two mechanisms are needed for providing a range of qualities of service to packets passing through a router or a tag switch. First, we need to classify packets into different classes. Second, we need to ensure that the handling of packets is such that the appropriate QOS characteristics (bandwidth, loss, etc.) are provided to each class.

Tag switching provides an easy way to mark packets as belonging to a particular class after they have been classified the first time. Initial classification would be done using information carried in the network layer or higher layer headers. A tag corresponding to the resultant class would then be applied to the packet. Tagged packets can then be efficiently handled by the tag switching routers in their path without needing to be reclassified. The actual packet scheduling and queueing is largely orthogonal - the key point here is that tag switching enables simple logic to be used to find the state that identifies how the packet should be scheduled.

The exact use of tag switching for QOS purposes depends a great deal on how QOS is deployed. If RSVP is used to request a certain QOS for a class of packets, then it would be necessary to allocate a tag corresponding to each RSVP session for which state is installed at a tag switch. This might be done by TDP or by extension of RSVP.

7. Tag switching migration strategies

Since tag switching is performed between a pair of adjacent tag switches, and since the tag binding information could be distributed on a pairwise basis, tag switching could be introduced in a fairly simple, incremental fashion. For example, once a pair of adjacent routers are converted into tag switches, each of the switches would tag packets destined to the other, thus enabling the other switch to use tag switching. Since tag switches use the same routing protocols as routers, the introduction of tag switches has no impact on routers. In fact, a tag switch connected to a router acts just as a router from the router’s perspective.

As more and more routers are upgraded to enable tag switching, the scope of functionality provided by tag switching widens. For example, once all the routers within a domain are upgraded to support tag switching, in becomes possible to start using the hierarchy of routing knowledge function.

8. Summary

In this document we described the tag switching technology. Tag switching is not constrained to a particular Network Layer protocol - it is a multiprotocol solution. The forwarding component of tag switching is simple enough to facilitate high performance forwarding, and may be implemented on high performance forwarding hardware such as ATM switches. The control component is flexible enough to support a wide variety of routing functions, such as destination-based routing, multicast routing, hierarchy of routing knowledge, and explicitly defined routes. By allowing a wide range of forwarding granularities that could be associated with a tag, we provide both scalable and functionally rich routing. A combination of a wide range of forwarding granularities and the ability to evolve the control component fairly independently from the forwarding component results in a solution that enables graceful introduction of new routing functionality to meet the demands of a rapidly evolving computer networking environment.

9. Security Considerations

Security issues are not discussed in this memo.

10. Intellectual Property Considerations

Cisco Systems may seek patent or other intellectual property protection for some or all of the technologies disclosed in this document. If any standards arising from this document are or become protected by one or more patents assigned to Cisco Systems, Cisco intends to disclose those patents and license them on reasonable and non-discriminatory terms.

11. Acknowledgments

Significant contributions to this work have been made by Anthony Alles, Fred Baker, Paul Doolan, Dino Farinacci, Guy Fedorkow, Jeremy Lawrence, Arthur Lin, Morgan Littlewood, Keith McCloghrie, and Dan Tappan.

12. Authors' Addresses

Yakov Rekhter
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134

EMail: yakov@cisco.com

Bruce Davie
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824

EMail: bsd@cisco.com

Dave Katz
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134

EMail: dkatz@cisco.com

Eric Rosen
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824

EMail: erosen@cisco.com

George Swallow
Cisco Systems, Inc.

250 Apollo Drive
Chelmsford, MA, 01824

EMail: swallow@cisco.com