Network Working Group
Request for Comments: 2917
Category: Informational

K. Muthukrishnan
Lucent Technologies
A. Malis
Vivace Networks, Inc.

September 2000

A Core MPLS IP VPN Architecture

Status of this Memo

Copyright Notice

Abstract

This memo presents an approach for building core Virtual Private Network (VPN) services in a service provider's MPLS backbone. This approach uses Multiprotocol Label Switching (MPLS) running in the backbone to provide premium services in addition to best effort services. The central vision is for the service provider to provide a virtual router service to their customers. The keystones of this architecture are ease of configuration, user security, network security, dynamic neighbor discovery, scaling and the use of existing routing protocols as they exist today without any modifications.

1. Acronyms

ARP	Address Resolution Protocol
CE	Customer Edge router
LSP	Label Switched Path
PNA	Private Network Administrator
SLA	Service Level Agreement
SP	Service Provider
SPED	Service Provider Edge Device
SPNA	SP Network Administrator
VMA	VPN Multicast Address
VPNID	VPN Identifier
VR	Virtual Router
VRC	Virtual Router Console

2. Introduction

This memo describes an approach for building IP VPN services out of the backbone of the SP's network. Broadly speaking, two possible approaches present themselves: the overlay model and the virtual router approach. The overlay model is based on overloading some semantic(s) of existing routing protocols to carry reachability information. In this document, we focus on the virtual router service.

The approach presented here does not depend on any modifications of any existing routing protocols. Neighbor discovery is aided by the use of an emulated LAN and is achieved by the use of ARP. This memo makes a concerted effort to draw the line between the SP and the PNA: the SP owns and manages layer 1 and layer 2 services while layer 3 services belong to and are manageable by the PNA. By the provisioning of fully logically independent routing domains, the PNA has been given the flexibility to use private and unregistered addresses. Due to the use of private LSPs and the use of VPNID encapsulation using label stacks over shared LSPs, data security is not an issue.

The approach espoused in this memo differs from that described in RFC 2547 [Rosen1] in that no specific routing protocol has been overloaded to carry VPN routes. RFC 2547 specifies a way to modify BGP to carry VPN unicast routes across the SP's backbone. To carry multicast routes, further architectural work will be necessary.

3. Virtual Routers

A virtual router is a collection of threads, either static or dynamic, in a routing device, that provides routing and forwarding services much like physical routers. A virtual router need not be a separate operating system process (although it could be); it simply has to provide the illusion that a dedicated router is available to satisfy the needs of the network(s) to which it is connected. A virtual router, like its physical counterpart, is an element in a routing domain. The other routers in this domain could be physical or virtual routers themselves. Given that the virtual router connects to a specific (logically discrete) routing domain and that a physical router can support multiple virtual routers, it follows that a physical router supports multiple (logically discreet) routing domains.

From the user (VPN customer) standpoint, it is imperative that the virtual router be as equivalent to a physical router as possible. In other words, with very minor and very few exceptions, the virtual router should appear for all purposes (configuration, management, monitoring and troubleshooting) like a dedicated physical router. The

main motivation behind this requirement is to avoid upgrading or re-configuring the large installed base of routers and to avoid retraining of network administrators.

The aspects of a router that a virtual router needs to emulate are:

1. Configuration of any combination of routing protocols

2. Monitoring of the network

3. Troubleshooting.

Every VPN has a logically independent routing domain. This enhances the SP's ability to offer a fully flexible virtual router service that can fully serve the SP's customer without requiring physical per-VPN routers. This means that the SP's "hardware" investments, namely routers and links between them, can be re-used by multiple customers.

4. Objectives

1. Easy, scalable configuration of VPN endpoints in the service provider network. At most, one piece of configuration should be necessary when a CE is added.

2. No use of SP resources that are globally unique and hard to get such as IP addresses and subnets.

3. Dynamic discovery of VRs (Virtual Routers) in the SP's cloud. This is an optional, but extremely valuable "keep it simple" goal.

4. Virtual Routers should be fully configurable and monitorable by the VPN network administrator. This provides the PNA with the flexibility to either configure the VPN themselves or outsource configuration tasks to the SP.

5. Quality of data forwarding should be configurable on a VPN-by-VPN basis. This should translate to continuous (but perhaps discrete) grades of service. Some examples include best effort, dedicated bandwidth, QOS, and policy based forwarding services.

6. Differentiated services should be configurable on a VPN-by-VPN basis, perhaps based on LSPs set up for exclusive use for forwarding data traffic in the VPN.

7. Security of internet routers extended to virtual routers. This means that the virtual router's data forwarding and routing functions should be as secure as a dedicated, private physical router. There should be no unintended leak of information (user data and reachability information) from one routing domain to another.

8. Specific routing protocols should not be mandated between virtual routers. This is critical to ensuring the VPN customer can setup the network and policies as the customer sees fit. For example, some protocols are strong in filtering, while others are strong in traffic engineering. The VPN customer might want to exploit both to achieve "best of breed" network quality.

9. No special extensions to existing routing protocols such as BGP, RIP, OSPF, ISIS etc. This is critical to allowing the future addition of other services such as NHRP and multicast. In addition, as advances and addenda are made to existing protocols (such as traffic engineering extensions to ISIS and OSPF), they can be easily incorporated into the VPN implementation.

5. Architectural Requirements

The service provider network must run some form of multicast routing to all nodes that will have VPN connections and to nodes that must forward multicast datagrams for virtual router discovery. A specific multicast routing protocol is not mandated. An SP may run MOSPF or DVMRP or any other protocol.

6. Architectural Outline

1. Every VPN is assigned a VPNID which is unique within the SP's network. This identifier unambiguously identifies the VPN with which a packet or connection is associated. The VPNID of zero is reserved; it is associated with and represents the public internet. It is recommended, but not required that these VPN identifiers will be compliant with RFC 2685 [Fox].

2. The VPN service is offered in the form of a Virtual Router service. These VRs reside in the SPED and are as such confined to the edge of the SP's cloud. The VRs will use the SP's network for data and control packet forwarding but are otherwise invisible outside the SPEDs.

3. The "size" of the VR contracted to the VPN in a given SPED is expressed by the quantity of IP resources such as routing interfaces, route filters, routing entries etc. This is entirely under the control of the SP and provides the fine granularity

that the SP requires to offer virtually infinite grades of VR service on a per-SPED level. [Example: one SPED may be the aggregating point (say headquarters of the corporation) for a given VPN and a number of other SPEDs may be access points (branch offices). In this case, the SPED connected to the headquarters may be contracted to provide a large VR while the SPEDs connected to the branch offices may house small, perhaps stub VRs]. This provision also allows the SP to design the network with an end goal of distributing the load among the routers in the network.

4. One indicator of the VPN size is the number of SPEDs in the SP's network that have connections to CPE routers in that VPN. In this respect, a VPN with many sites that need to be connected is a "large" VPN whereas one with a few sites is a "small" VPN. Also, it is conceivable that a VPN grows or shrinks in size over time. VPNs may even merge due to corporate mergers, acquisitions and partnering agreements. These changes are easy to accommodate in this architecture, as globally unique IP resources do not have to be dedicated or assigned to VPNs. The number of SPEDs is not limited by any artificial configuration limits.

5. The SP owns and manages Layer 1 and Layer 2 entities. To be specific, the SP controls physical switches or routers, physical links, logical layer 2 connections (such as DLCI in Frame Relay and VPI/VCI in ATM) and LSPs (and their assignment to specific VPNs). In the context of VPNs, it is the SP's responsibility to contract and assign layer 2 entities to specific VPNs.

6. Layer 3 entities belong to and are manageable by the PNA. Examples of these entities include IP interfaces, choice of dynamic routing protocols or static routes, and routing interfaces. Note that although Layer 3 configuration logically falls under the PNA's area of responsibility, it is not necessary for the PNA to execute it. It is quite viable for the PNA to outsource the IP administration of the virtual routers to the Service Provider. Regardless of who assumes responsibility for configuration and monitoring, this approach provides a full routing domain view to the PNA and empowers the PNA to design the network to achieve intranet, extranet and traffic engineering goals.

7. The VPNs can be managed as if physical routers rather than VRs were deployed. Therefore, management may be performed using SNMP or other similar methods or directly at the VR console (VRC).

8. Industry-standard troubleshooting tools such as ’ping,’ ’traceroute,’ in a routing domain domain comprised exclusively of dedicated physical routers. Therefore, monitoring and .bp troubleshooting may be performed using SNMP or similar methods, but may also include the use of these standard tools. Again, the VRC may be used for these purposes just like any physical router.

9. Since the VRC is visible to the user, router specific security checks need to be put in place to make sure the VPN user is allowed access to Layer 3 resources in that VPN only and is disallowed from accessing physical resources in the router. Most routers achieve this through the use of database views.

10. The VRC is available to the SP as well. If configuration and monitoring has been outsourced to the SP, the SP may use the VRC to accomplish these tasks as if it were the PNA.

11. The VRs in the SPEDs form the VPN in the SP’s network. Together, they represent a virtual routing domain. They dynamically discover each other by utilizing an emulated LAN resident in the SP’s network.

Each VPN in the SP’s network is assigned one and only one multicast address. This address is chosen from the administratively scoped range (239.192/14) [Meyer] and the only requirement is that the multicast address can be uniquely mapped to a specific VPN. This is easily automated by routers by the use of a simple function to unambiguously map a VPNid to the multicast address. Subscription to this multicast address allows a VR to discover and be discovered by other VRs. It is important to note that the multicast address does not have to be configured.

12. Data forwarding may be done in one of several ways:

    1. An LSP with best-effort characteristics that all VPNS can use.

    2. An LSP dedicated to a VPN and traffic engineered by the VPN customer.

    3. A private LSP with differentiated characteristics.

    4. Policy based forwarding on a dedicated L2 Virtual Circuit

The choice of the preferred method is negotiable between the SP and the VPN customer, perhaps constituting part of the SLA between them. This allows the SP to offer different grades of service to different VPN customers.

Of course, hop-by-hop forwarding is also available to forward routing packets and to forward user data packets during periods of LSP establishment and failure.

13. This approach does not mandate that separate operating system tasks for each of the routing protocols be run for each VR that the SPED houses. Specific implementations may be tailored to the particular SPED in use. Maintaining separate routing databases and forwarding tables, one per VR, is one way to get the highest performance for a given SPED.

7. Scalable Configuration

A typical VPN is expected to have 100s to 1000s of endpoints within the SP cloud. Therefore, configuration should scale (at most) linearly with the number of end points. To be specific, the administrator should have to add a couple of configuration items when a new customer site joins the set of VRs constituting a specific VPN. Anything worse will make this task too daunting for the service provider. In this architecture, all that the service provider needs to allocate and configure is the ingress/egress physical link (e.g. Frame Relay DLCI or ATM VPI/VCI) and the virtual connection between the VR and the emulated LAN.

8. Dynamic Neighbor Discovery

The VRs in a given VPN reside in a number of SPEDs in the network. These VRs need to learn about each other and be connected.

One way to do this is to require the manual configuration of neighbors. As an example, when a new site is added to a VPN, this would require the configuration of all the other VRs as neighbors. This is obviously not scalable from a configuration and network resource standpoint.

The need then arises to allow these VRs to dynamically discover each other. Neighbor discovery is facilitated by providing each VPN with a limited emulated LAN. This emulated LAN is used in several ways:

1. Address resolution uses this LAN to resolve next-hop (private) IP addresses associated with the other VRs.

2. Routing protocols such as RIP and OSPF use this limited emulated LAN for neighbor discovery and to send routing updates.

The per-VPN LAN is emulated using an IP multicast address. In the interest of conserving public address space and because this multicast address needs to be visible only in the SP network space,

we would use an address from the Organizationally scoped multicast addresses (239.192/14) as described in [Meyer]. Each VPN is allocated an address from this range. To completely eliminate configuration in this regard, this address is computed from the VPNID.

9. VPN IP Domain Configuration

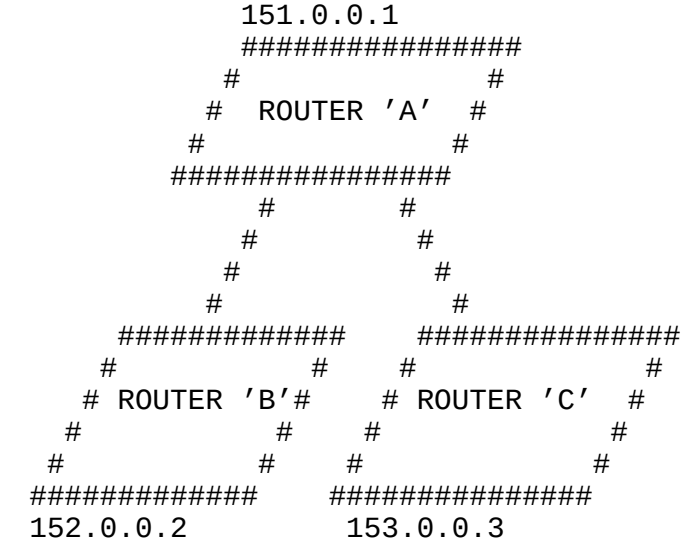

Figure 1 'Physical Routing Domain'

The physical domain in the SP's network is shown in the above figure. In this network, physical routers A, B and C are connected together. Each of the routers has a 'public' IP address assigned to it. These addresses uniquely identify each of the routers in the SP's network.

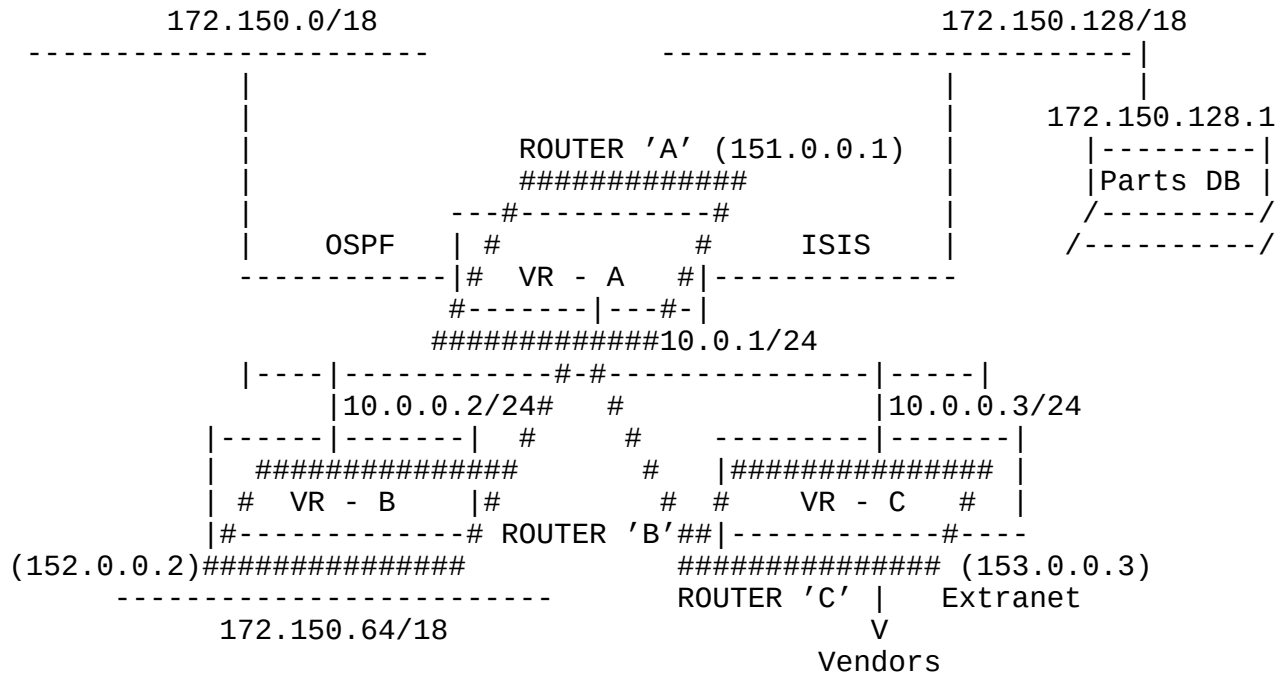

Figure 2 'Virtual Routing Domain'

Each Virtual Router is configurable by the PNA as though it were a private physical router. Of course, the SP limits the resources that this Virtual Router may consume on a SPED-by-SPED basis. Each VPN has a number of physical connections (to CPE routers) and a number of logical connections (to the emulated LAN). Each connection is IP-capable and can be configured to utilize any combination of the standard routing protocols and routing policies to achieve specific corporate network goals.

To illustrate, in Figure 1, 3 VRs reside on 3 SPEDs in VPN 1. Router 'A' houses VR-A, router 'B' houses VR-B and router 'C' houses VR-C. VR-C and VR-B have a physical connection to CPE equipment, while VR-A has 2 physical connections. Each of the VRs has a fully IP-capable logical connection to the emulated LAN. VR-A has the (physical) connections to the headquarters of the company and runs OSPF over those connections. Therefore, it can route packets to 172.150.0/18 and 172.150.128/18. VR-B runs RIP in the branch office (over the physical connection) and uses RIP (over the logical connection) to export 172.150.64/18 to VR-A. VR-A advertises a default route to VR-B over the logical connection. Vendors use VR-C as the extranet connection to connect to the parts database at 172.150.128.1. Hence, VR-C advertises a default route to VR-A over the logical connection. VR-A exports only 175.150.128.1 to VR-C. This keeps the rest of the corporate network from a security problem.

The network administrator will configure the following:

1. OSPF connections to the 172.150.0/18 and 172.150.128/18 network in VR-A.

2. RIP connections to VR-B and VR-C on VR-A.

3. Route policies on VR-A to advertise only the default route to VR-B.

4. Route policies on VR-A to advertise only 172.159.128.1 to VR-C.

5. RIP on VR-B to VR-A.

6. RIP on VR-C to advertise a default route to VR-A.

10. Neighbor Discovery Example

In Figure #1, the SPED that houses VR-A (SPED-A) uses a public address of 150.0.0.1/24, SPED-B uses 150.0.0.2/24 and SPED-C uses 150.0.0.3/24. As noted, the connection between the VRs is via an emulated LAN. For interface addresses on the emulated LAN connection, VR-A uses 10.0.0.1/24, VR-B uses 10.0.0.2/24 and VR-C uses 10.0.0.3/24.

Let’s take the case of VR-A sending a packet to VR-B. To get VR-B’s address (SPED-B’s address), VR-A sends an ARP request packet with the address of VR-B (10.0.0.2) as the logical address. The source logical address is 10.0.0.1 and the hardware address is 151.0.0.1. This ARP request is encapsulated in this VPN’s multicast address and sent out. SPED B and SPED-C receive a copy of the packet. SPED-B recognizes 10.0.0.2 in the context of VPN 1 and responds with 152.0.0.2 as the "hardware" address. This response is sent to the VPN multicast address to promote the use of promiscuous ARP and the resulting decrease in network traffic.

Manual configuration would be necessary if neighbor discovery were not used. In this example, VR-A would be configured with a static ARP entry for VR-B’s logical address (10.0.0.1) with the "hardware" address set to 152.0.0.2.

11. Forwarding

As mentioned in the architectural outline, data forwarding may be done in one of several ways. In all techniques except the Hop-by-Hop technique for forwarding routing/control packets, the actual method

is configurable. At the high end, policy based forwarding for quick service and at the other end best effort forwarding using public LSP is used. The order of forwarding preference is as follows:

1. Policy based forwarding.

2. Optionally configured private LSP.

3. Best-effort public LSP.

11.1 Private LSP

This LSP is optionally configured on a per-VPN basis. This LSP is usually associated with non-zero bandwidth reservation and/or a specific differentiated service or QOS class. If this LSP is available, it is used for user data and for VPN private control data forwarding.

11.2 Best Effort Public LSP

VPN data packets are forwarded using this LSP if a private LSP with specified bandwidth and/or QOS characteristics is either not configured or not presently available. The LSP used is the one destined for the egress router in VPN 0. The VPNID in the shim header is used to de-multiplex data packets from various VPNs at the egress router.

12. Differentiated Services

Configuring private LSPs for VPNs allows the SP to offer differentiated services to paying customers. These private LSPs could be associated with any available L2 QOS class or Diff-Serv codepoints. In a VPN, multiple private LSPs with different service classes could be configured with flow profiles for sorting the packets among the LSPs. This feature, together with the ability to size the virtual routers, allows the SP to offer truly differentiated services to the VPN customer.

13. Security Considerations

13.1 Routing Security

The use of standard routing protocols such as OSPF and BGP in their unmodified form means that all the encryption and security methods (such as MD5 authentication of neighbors) are fully available in VRs. Making sure that routes are not accidentally leaked from one VPN to another is an implementation issue. One way to achieve this is to maintain separate routing and forwarding databases.

13.2 Data Security

This allows the SP to assure the VPN customer that data packets in one VPN never have the opportunity to wander into another. From a routing standpoint, this could be achieved by maintaining separate routing databases for each virtual router. From a data forwarding standpoint, the use of label stacks in the case of shared LSPs [Rosen2] [Callon] or the use of private LSPs guarantees data privacy. Packet filters may also be configured to help ease the problem.

13.3 Configuration Security

Virtual routers appear as physical routers to the PNA. This means that they may be configured by the PNA to achieve connectivity between offices of a corporation. Obviously, the SP has to guarantee that the PNA and the PNA's designees are the only ones who have access to the VRs on the SPEDs the private network has connections to. Since the virtual router console is functionally equivalent to a physical router, all of the authentication methods available on a physical console such as password, RADIUS, etc. are available to the PNA.

13.4 Physical Network Security

When a PNA logs in to a SPED to configure or monitor the VPN, the PNA is logged into the VR for the VPN. The PNA has only layer 3 configuration and monitoring privileges for the VR. Specifically, the PNA has no configuration privileges for the physical network. This provides the guarantee to the SP that a VPN administrator will not be able to inadvertently or otherwise adversely affect the SP's network.

14. Virtual Router Monitoring

All of the router monitoring features available on a physical router are available on the virtual router. This includes utilities such as "ping" and "traceroute". In addition, the ability to display private routing tables, link state databases, etc. are available.

15. Performance Considerations

For the purposes of discussing performance and scaling issues, today's routers can be split into two planes: the routing (control) plane and the forwarding plane.

In looking at the routing plane, most modern-day routing protocols use some form of optimized calculation methodologies to calculate the shortest path(s) to end stations. For instance, OSPF and ISIS use the Djikstra algorithm while BGP uses the "Decision Process". These

algorithms are based on parsing the routing database and computing the best paths to end stations. The performance characteristics of any of these algorithms is based on either topological characteristics (ISIS and OSPF) or the number of ASs in the path to the destinations (BGP). But it is important to note that the overhead in setting up and beginning these calculations is very little for most any modern day router. This is because, although we refer to routing calculation input as "databases", these are memory resident data structures.

Therefore, the following conclusions can be drawn:

1. Beginning a routing calculation for a routing domain is nothing more than setting up some registers to point to the right database objects.

2. Based on 1, the performance of a given algorithm is not significantly worsened by the overhead required to set it up.

3. Based on 2, it follows that, when a number of routing calculations for a number of virtual routers has to be performed by a physical router, the complexity of the resulting routing calculation is nothing more than the sum of the complexities of the routing calculations of the individual virtual routers.

4. Based on 3, it follows that whether an overlay model is used or a virtual routing model is employed, the performance characteristics of a router are dependent purely on its hardware capabilities and the choice of data structures and algorithms.

To illustrate, let’s say a physical router houses N VPNs, all running some routing protocol say RP. Let’s also suppose that the average performance of RP’s routing calculation algorithm is f(X,Y) where x and y are parameters that determine performance of the algorithm for that routing protocol. As an example, for Djikstra algorithm users such as OSPF, X could be the number of nodes in the area while Y could be the number of links. The performance of an arbitrary VPN n is f (Xn, Yn). The performance of the (physical) router is the sum of f(Xi, Yi) for all values of i in 0 <= i <= N. This conclusion is independent of the chosen VPN approach (virtual router or overlay model).

In the usual case, the forwarding plane has two inputs: the forwarding table and the packet header. The main performance parameter is the lookup algorithm. The very best performance one can get for a IP routing table lookup is by organizing the table as some form of a tree and use binary search methods to do the actual lookup. The performance of this algorithm is O(log n).

Hence, as long as the virtual routers' routing tables are distinct from each other, the lookup cost is constant for finding the routing table and O(log n) to find the entry. This is no worse or different from any router and no different from a router that employs overlay techniques to deliver VPN services. However, when the overlay router utilizes integration of multiple VPNs' routing tables, the performance is O(log m*n) where 'm' is the number of VPNs that the routing table holds routes for.

16. Acknowledgements

The authors wish to thank Dave Ryan, Lucent Technologies for his invaluable in-depth review of this version of this memo.

17. References

[Callon] Callon R., et al., "A Framework for Multiprotocol Label Switching", Work in Progress.

[Fox] Fox, B. and B. Gleeson,"Virtual Private Networks Identifier", RFC 2685, September 1999.

[Meyer] Meyer, D., "Administratively Scoped IP Multicast", RFC 2365, July 1998.

[Rosen1] Rosen, E. and Y. Rekhter, "BGP/MPLS VPNs", RFC 2547, March 1999.

[Rosen2] Rosen E., Viswanathan, A. and R. Callon, "Multiprotocol Label Switching Architecture", Work in Progress.

18. Authors' Addresses

Karthik Muthukrishnan
Lucent Technologies
1 Robbins Road
Westford, MA 01886

Phone: (978) 952-1368
EMail: mkarthik@lucent.com

Andrew Malis
Vivace Networks, Inc.
2730 Orchard Parkway
San Jose, CA 95134

Phone: (408) 383-7223
EMail: Andy.Malis@vivacenetworks.com

19. Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.