Network Working Group
Internet Draft

Expires: November 2003
draft-ietf-l2vpn-vpls-bgp-00.txt

K. Kompella (Juniper)
Y. Rekhter (Juniper)
V. Kompella (TiMetra)
J. Achirica (Telefonica)
L. Andersson (Utfors)
G. Heron (PacketExchange)
S. Khandekar (TiMetra)
M. Lasserre (Riverstone)
P. Lin (Yipes)
P. Menezes (Terabeam)
A. Moranganti (Appian)
H. Ould-Brahim (Nortel)
S. Yeong-il (Korea Tel)

Virtual Private LAN Service

Status of this Memo

Abstract

Virtual Private LAN Service (VPLS), also known as Transparent LAN Service, and Virtual Private Switched Network service, is a useful Service Provider offering. The service offered is a Layer 2 VPN; however, in the case of VPLS, the customers in the VPN are connected by a multipoint network, in contrast to the usual Layer 2 VPNs, which are point-to-point in nature.

This document describes the functions required to offer VPLS, and proposes a mechanism for signaling a VPLS, as well as for forwarding VPLS frames across a packet switched network.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [1].

The terms P (Provider router, VPN-unaware), PE (Provider Edge router), VE (VPLS Edge device), CE (Customer Edge device), etc. are defined in [2].

1. Introduction

Virtual Private LAN Service (VPLS), also known as Transparent LAN Service, and Virtual Private Switched Network service, is a useful service offering. A Virtual Private LAN appears in (almost) all respects as a LAN to customers of a Service Provider. However, in a VPLS, the customers are not all connected to a single LAN; the customers may be spread across a metro or wide area. In essence, a VPLS glues several individual LANs across a metro area to appear and function as a single LAN [3].

This document describes the functions needed to offer VPLS, and goes on to propose a mechanism for signaling a VPLS, as well as a

mechanism for transport of VPLS frames over tunnels across a packet switched network. The signaling mechanism is taken from [4]; BGP is used as the control plane protocol. This document also discusses deployment options, in particular, the notion of decoupling functions across devices.

Alternative approaches include: [4], which allows one to build a Layer 2 VPN with Ethernet as the interconnect; and [5], which allows one to set up an Ethernet connection across a packet switched network. Both of these, however, offer point-to-point Ethernet services. What distinguishes VPLS from the above two is that a VPLS offers a multipoint service. A mechanism for setting up pseudowires for VPLS using LDP is defined in [6].

1.1. Scope of this Document

This document has four major parts: defining a VPLS functional model; defining a control plane for setting up VPLS; defining the data plane for VPLS (encapsulation and forwarding of data); and defining various deployment scenarios.

The functional model underlying VPLS is laid out in section 2. This describes the service being offered, the network components that interact to provide the service, and at a high level their interactions.

The control plane proposed here (section 3) uses BGP to establish VPLS service, i.e., for autodiscovery of VPLS members and for the setup and teardown of the pseudowires that constitute a given VPLS. Section 3 also describes how a VPLS that spans Autonomous System boundaries is set up. Using BGP as the control plane for VPNs is not new (see [4], [7] and [8]): what is described here is identical to what is in [4], which itself is based on the mechanisms proposed in [7].

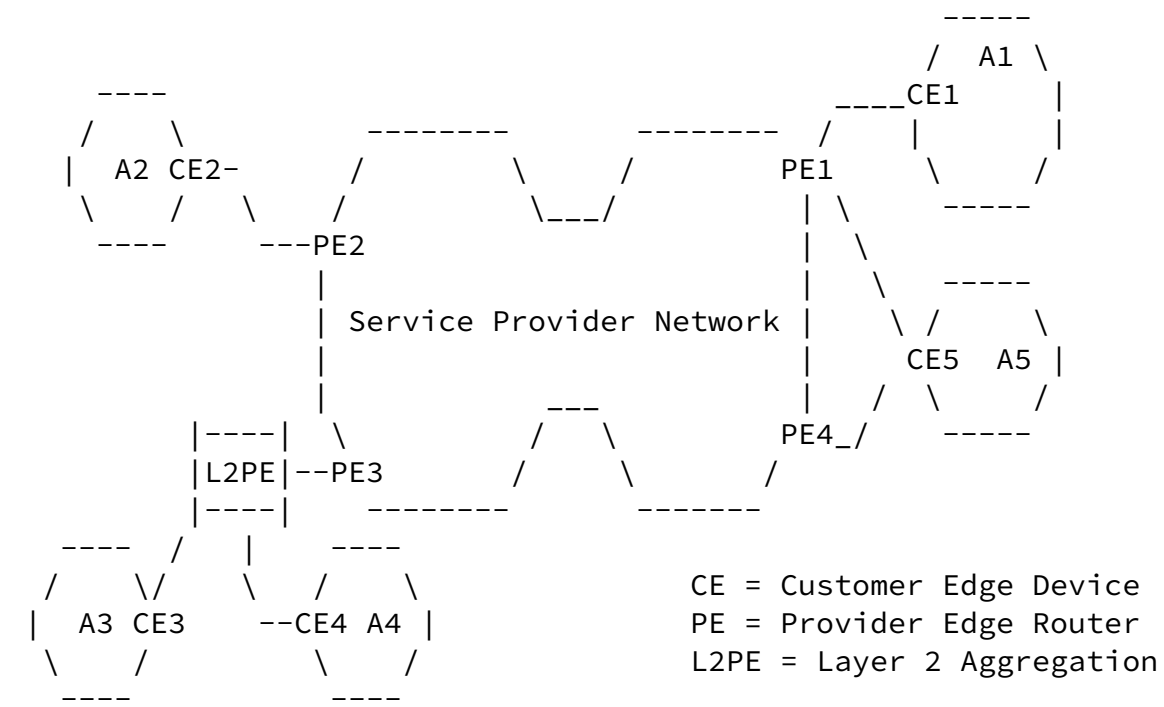
The forwarding plane and the actions that a participating PE must take is described in section 4.

In section 5, the notion of 'decoupled' operation is defined, and the interaction of decoupled and non-decoupled PEs is described. Decoupling allows for more flexible deployment of VPLS.

2. Functional Model

This will be described with reference to Figure 1.

Figure 1: Example of a VPLS



2.1. Terminology

(NOTE: the terminology here has not quite been fully harmonized with the terminology in the VPN terminology document [2] or the PWE3

framework document; this will be done in a later version.)

The terminology of [4] is used, with the addition of "L2PE", a Layer 2 PE device used for Layer 2 aggregation. An L2PE is owned and operated by the Service Provider (as is the PE). PE and L2PE devices are "VPLS-aware", which means that they know that a VPLS service is being offered. We will call these VPLS edge devices, which could be either a PE or an L2PE, a VE.

In contrast, the CE device (which may be owned and operated by either the SP or the customer) is VPLS-unaware; as far as the CE is concerned, it is connected to the other CEs in the VPLS via a Layer 2 switched network. This means that there should be no changes to a CE device, either to the hardware or the software, in order to offer VPLS.

A CE device may be connected to a PE or an L2PE via Layer 2 switches that are VPLS-unaware. From a VPLS point of view, such Layer 2 switches are invisible, and hence will not be discussed further.

Furthermore, an L2PE may be connected to a PE via Layer 2 and Layer 3 devices; this will be discussed further in a later section.

The term "demultiplexor" refers to an identifier in a data packet that identifies both the VPLS to which the packet belongs as well as the ingress PE. In this document, the demultiplexor is an MPLS label.

The term "VPLS" will refer to the service as well as a particular instantiation of the service (i.e., an emulated LAN); it should be clear from the context which usage is intended.

2.2. Assumptions

The Service Provider Network is a packet switched network. The PEs are assumed to be full meshed with tunnels over which packets that belong to a service (such as VPLS) are encapsulated and forwarded. These tunnels can be IP tunnels, such as GRE, or MPLS tunnels, established by RSVP-TE or LDP. These tunnels are established independently of the services offered over them; the signaling and establishment of these tunnels are not discussed in this document.

"Flooding" and MAC address "learning" (see section 4) are an integral part of VPLS. However, these activities are private to an SP device, i.e., in the VPLS described below, no SP device requests another SP device to flood packets or learn MAC addresses on its behalf.

All the PEs participating in a VPLS are assumed to be fully meshed, i.e., every (ingress) PE can send a VPLS packet to the egress PE(s) directly, without the need for an intermediate PE. This assumption reduces (but does not eliminate) the need to run Spanning Tree Protocol among the PEs.

2.3. Interactions

VPLS is a successful "LAN Service" if CE devices that belong to VPLS V can interact through the SP network as if they were connected by a LAN. VPLS is "private" if CE devices that belong to different VPLSs cannot interact. VPLS is "virtual" if multiple VPLSs can be offered over a common packet switched network.

PE devices interact to "discover" who all participate in the same VPLS (i.e., are attached to CE devices that belong to the same VPLS), and to exchange demultiplexors. These interactions are control-driven, not data-driven.

L2PEs interact with PEs to establish connections with remote PEs or L2PEs in the same VPLS. Again, this interaction is control-driven.

3. Control Plane

There are two primary functions of the VPLS control plane: autodiscovery, and setup and teardown of the pseudowires that constitute the VPLS, often called signaling. The first two subsections describe these functions. The last subsection describes the setting up of pseudowires that span Autonomous Systems.

3.1. Autodiscovery

Autodiscovery refers to the process of finding all the PEs that participate in a given VPLS. A PE can either be configured with the identities of all the other PEs in a given VPLS, or the PE can autodiscover the other PEs.

The former approach is fairly configuration-intensive, especially since it is required (in this and other VPLS approaches) that the PEs participating in a given VPLS are fully meshed (i.e., every pair of PEs in a given VPLS establish pseudowires to each other). Furthermore, when the topology of a VPLS changes (i.e., a PE is added to, or removed from the VPLS), the VPLS configuration on all PEs in that VPLS must be changed.

In the autodiscovery approach, each PE "discovers" which other PEs are part of a given VPLS by means of some protocol. In this approach, each PE's configuration consists only of the identity of the VPLS that each customer belongs to, not the identity of every other PE in that VPLS. Moreover, when the topology of a VPLS changes, only the affected PE's configuration changes; other PEs automatically find out about the change and adapt.

3.1.1. Functions

A PE that participates in a given VPLS V must be able to tell all other PEs in VPLS V that it is also a member of V. A PE must also have a means of declaring that it no longer participates in a VPLS. To do both of these, the PE must have a means of identifying a VPLS and a means by which to communicate to all other PEs.

L2PE devices also need to know what constitutes a given VPLS; however, they don't need the same level of detail. The PE (or PEs) to which an L2PE is connected gives the L2PE an abstraction of the VPLS; this is described in section 5. One protocol mechanism to achieve this is described in [9].

3.1.2. Protocol Specification

The specific mechanism for autodiscovery described here, borrowed essentially unchanged from [4] and [7], uses BGP extended communities [10] to identify a VPLS. This mechanism is described more generically in [8]. The specific extended community used is the Route Target, whose format is described in [10]. The semantics of the use of Route Targets is described in [7]; their use in VPLS is

identical.

As it has been assumed that VPLSs are fully meshed, a single Route Target RT suffices for a given VPLS V, and in effect that RT is the identifier for VPLS V.

A PE announces (typically via I-BGP) that it belongs to VPLS V by annotating its NLRIs for V (see next subsection) with Route Target RT, and acts on this by accepting NLRIs from other PEs that have Route Target RT. A PE announces that it no longer participates in V by withdrawing all NLRIs that it had advertised with Route Target RT.

3.2. Signaling

Once discovery is done, each pair of PEs in a VPLS must be able to establish (and tear down) pseudowires to each other, i.e., exchange (and withdraw) demultiplexors. This process is known as signaling. Signaling is also used to initiate "relearning", and to transmit certain characteristics of the PE regarding a given VPLS.

Recall that a demultiplexor is used to distinguish among several different streams of traffic carried over a tunnel, each stream possibly representing a different service. In the case of VPLS, the demultiplexor not only says to which specific VPLS a packet belongs, but also identifies the ingress PE. The former information is used for forwarding the packet; the latter information is used for learning MAC addresses. The demultiplexor described here is an MPLS label, even though the PE-to-PE tunnels may not be MPLS tunnels.

3.2.1. Setup and Teardown

A BGP NLRI, the VPLS NLRI, is used to exchange demultiplexors, using the mechanism described in [4].

A PE advertises a VPLS NLRI for each VPLS that it participates in. If the PE is doing learning and flooding, i.e., it is the VE, it announces a single set of VPLS NLRIs for each VPLS that it is in. If the PE is connected to several L2PEs, it announces one set of VPLS NLRIs for each L2PE. A hybrid scheme is also possible, where the PE learns MAC addresses on some interfaces (over which it is directly

connected to CEs) and delegates learning on other interfaces (over

which it is connected to L2PEs). In this case, the PE would announce one set of VPLS NLRIs for each L2PE that has customer ports in a given VPLS, and one set for itself, if it has customer ports in that VPLS.

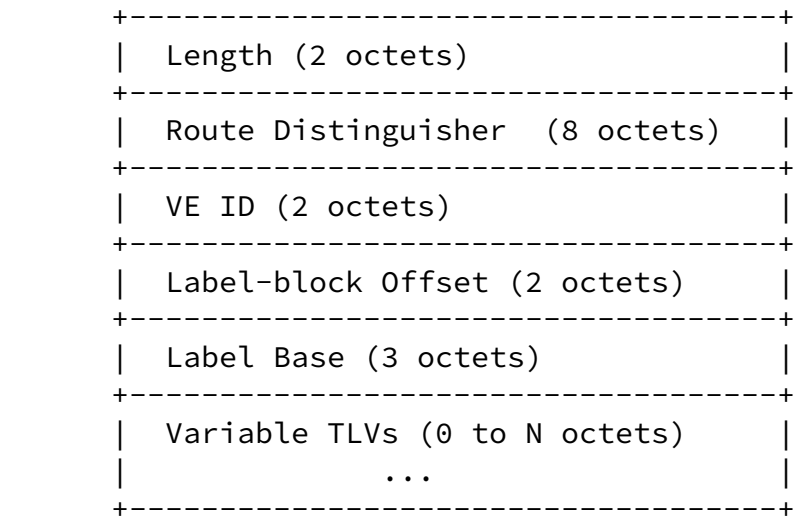
Each set of NLRIs defines the demultiplexors for a range of other PEs in the VPLS. Ideally, a single NLRI suffices for all PEs in a VPLS; however, there are cases (such as a newly added PE) where the pre-existing NLRI does not have enough labels. In such cases, advertising an additional NLRI for the same VPLS serves to add labels for the new PEs without disrupting service to the pre-existing PEs. If service disruption is acceptable (or when the PE restarts its BGP process), a PE MAY consider coalescing all NLRIs for a VPLS into a single NLRI.

If a PE X is part of VPLS V, and X receives a VPLS NLRI for V from PE Y that includes a demultiplexor that X can use, X sets up its ends of a pair of pseudowires between X and Y. X may also have to advertise a new NLRI for V that includes a demultiplexor that Y can use, if its pre-existing NLRI for V did not include a demultiplexor for Y.

If Y withdraws its NLRI for V that X was using, then X tears down its ends of the pseudowires between X and Y.

The format of the VPLS NLRI is given below; it is essentially identical to the L2 VPN NLRI [4]. The AFI and SAFI are the same as for the L2 VPN NLRI.

Figure 2: BGP NLRI for VPLS Information



3.2.2. Relearning MAC Addresses

Once a MAC address has been learned by VE A, VE A no longer needs to flood packets destined to that MAC address; instead VE A can send those packets directly to the VE "owning" that MAC address, say B. However, it is possible that a CE "move" from VE B to VE C; one example scenario is that the CE is dual-homed to VE B and C, and the link over which the CE is attached to VE B goes down. In this case, VE B may want to signal to other VEs in the VPLS that MAC addresses that they learned from VE B (for the given VPLS) are no longer valid. While aging timers will eventually enforce this, they may often be too slow. The Relearn Sequence Number (RSN) TLV will help speed up relearning.

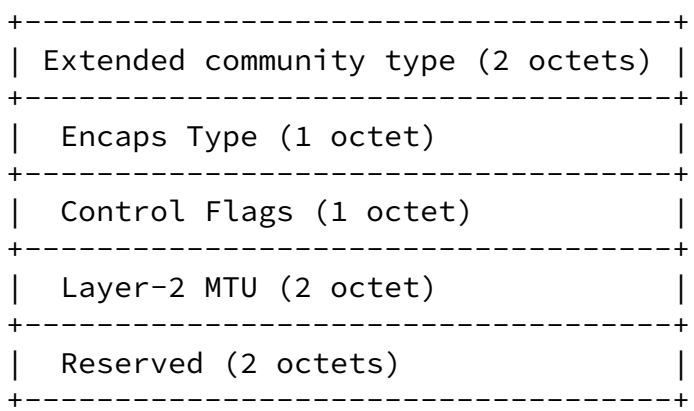
The RSN TLV is an optional TLV with Type TBD, Length 4 and Value a 32 bit RSN, a monotonically increasing unsigned number. When an RSN TLV is received, the RSN number is compared against the existing one for that VPLS and PE. If the new number is higher than the previous one, or no previous RSN exists, the PE SHOULD flush all existing MAC address to VC bindings for that VPLS and PE.

3.2.3. Signaling PE Capabilities

The Encaps Type and Control Flags are encoded in an extended attribute, just as in [4]. The community type remains the same.

There is a new Encaps Type for VPLS (TBD).

Figure 3: layer2-info extended community

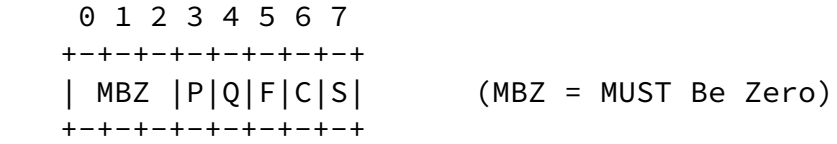


There are three new control flags, Q, F and P defined for VPLS. Q says whether qualified learning occurs (1) or not (0); F which says

whether the PE is capable of flooding (1) or not (0). P indicates that the PE will strip the outermost VLAN from the layer 2 customer

frame on ingress, and push a VLAN on egress.

Figure 4: Control Flags Bit Vector



3.3. Inter-Provider VPLS

As in [4] and [7], the above autodiscovery and signaling functions are typically announced via I-BGP. This assumes that all the sites in a VPLS are connected to PEs in a single Autonomous System (AS).

However, sites in a VPLS may occasionally connect to PEs in different ASes. This leads to two issues: a) there would not be an I-BGP connection between those PEs, so some means of signaling inter-AS is needed; and b) there may not be PE-to-PE tunnels between the ASes.

The former problem is solved in [7], Section 10. Three methods are suggested; of these, the last two Just Work for Inter-Provider VPLS. Method (b) requires an I-BGP peering between the PEs in AS1 and ASBR1 in AS1, an E-BGP peering between ASBR1 and ASBR2 in AS2, and I-BGP peerings between ASBR2 and the PEs in AS2. Method (c) requires a multi-hop E-BGP peering between the PEs (or preferably, a Route Reflector) in AS1 and the PEs (or Route Reflector) in AS2.

The latter is easy if the PE-to-PE tunnels are IP. If the tunnels are MPLS, labeled IPv4 distribution of PE loopback addresses by ASBRs (as described in part (c) of Section 10 of [7]) can be used to create PE-to-PE MPLS LSPs that traverses the ASes.

4. Data Plane

This section discusses two aspects of the data plane for PEs and

L2PEs implementing VPLS: encapsulation and forwarding.

4.1. Encapsulation

Ethernet frames received from CE devices are encapsulated for transmission over the packet switched network connecting the PEs. The encapsulation is as in [11], with one change: a PE that sets the P bit in the Control Flags strips the outermost VLAN from an Ethernet frame received from a CE before encapsulating it, and pushes a VLAN

onto a decapsulated frame before sending it to a CE.

4.2. Forwarding

Forwarding of VPLS packets is based on the interface over which the packet is received, which determines which VPLS the packet belongs to, and the destination MAC address. The former mapping is determined by configuration. The latter is the focus of this section.

4.2.1. MAC address learning

As was mentioned earlier, the key distinguishing feature of VPLS is that it is a multipoint service. This means that the entire Service Provider network should appear as a single logical learning bridge for each VPLS that the SP network supports. The logical ports for the SP "bridge" are the connections from the SP edge, be it a PE or an L2PE, to the CE. Just as a learning bridge learns MAC addresses on its ports, the SP bridge must learn MAC addresses at its VEs [3].

Learning consists of associating source MAC addresses of packets with the ports on which they arrive; call this association the Forwarding Information Base (FIB). The FIB is used for forwarding packets. For example, suppose the bridge receives a packet with source MAC address S on port P. If subsequently, the bridge receives a packet with destination MAC address S, it knows that it should send the packet out on port P.

There are two modes of learning: qualified and unqualified learning.

In qualified learning, the learning decisions at the VE are based on the customer ethernet packet's MAC address and VLAN tag, if one

exists. If no VLAN tag exists, the default VLAN is assumed. Effectively, within one VPLS, there are multiple logical FIBs, one for each customer VLAN tag identified in a customer packet.

In unqualified learning, learning is based on a customer ethernet packet's MAC address only. In other words, at any VE, there is only one FIB per VPLS.

Every VE must have at least one FIB for each VPLS that it participates in.

4.2.2. Flooding

When a bridge receives a packet to a destination that is not in its FIB, it floods the packet on all the other ports. Similarly, a VE will flood packets to an unknown destination to all other VEs in the VPLS.

In Figure 1 above, if CE2 sent an Ethernet frame to PE2, and the destination MAC address on the frame was not in PE2's FIB (for that VPLS), then PE2 would be responsible for flooding that frame to every other PE in the same VPLS. On receiving that frame, PE1 would be responsible for further flooding the frame to CE1 and CE5 (unless PE1 knew which CE "owned" that MAC address).

On the other hand, if PE3 received the frame, it could delegate further flooding of the frame to its L2PE. If PE3 was connected to 2 L2PEs, it would announce that it has two L2PEs. PE3 could either announce that it is incapable of flooding, in which case it would receive two frames, one for each L2PE, or it could announce that it is capable of flooding, in which case it would receive one copy of the frame, which it would then send to both L2PEs.

5. Deployment Scenarios

In deploying a network that supports VPLS, the SP must decide whether the VPLS-aware device closest to the customer (the VE) is an L2PE or a PE. The default case described in this document is that the VE is a PE. However, there are a number of reasons that the VE might be an L2PE, i.e., a device that does layer 2 functions such as MAC address learning and flooding, and some limited layer 3 functions such as communicating to its PE, but doesn't do full-fledged discovery and PE-to-PE signaling [12].

As both of these cases have benefits, one would like to be able to "mix and match" these scenarios. The signaling mechanism presented here allows this. PE1 may be directly connected to CE devices; PE2 may be connected to L2PEs that are connected to CEs; and PE3 may be connected directly to a customer over some interfaces and to L2PEs over others. All these PEs do discovery and signaling in the same manner. How they do learning and forwarding depends on whether or not there is an L2PE; however, this is a local matter, and is not signaled.

6. Normative References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997

[2] Andersson, L. and T. Madsen, "PPVPN Terminology", work in progress.

[3] Andersson, L. (Editor), "PPVPN L2 Framework", work in progress.

[4] Kompella, K., et al, "MPLS-based Layer 2 VPNs", work in progress.

[7] Rosen, E., et al, "BGP/MPLS VPNs", work in progress.

[10] Sangli, S., D. Tappan, and Y. Rekhter, "BGP Extended Communities Attribute", (work in progress).

[11] Martini, L., et al, "Encapsulation Methods for Transport of Layer 2 Frames Over IP and MPLS Networks", work in progress.

[12] Kompella, K., et al, "Decoupled TLS", work in progress.

7. Informative References

[5] Martini, L., et al, "Transport of Layer 2 Frames Over MPLS", work in progress.

[6] Kompella, V., et al, "Virtual Private Switched Network Services over an MPLS Network", work in progress.

[8] Ould-Brahim, H. et al, "Using BGP as an Auto-Discovery Mechanism for Network-based VPNs", work in progress.

[9] Shah, H. et al, "Signaling between PE and L2PE/MTU for Decoupled VPLS and Hierarchical VPLS", work in progress.

Security Considerations

To be filled in in a later version.

IANA Considerations

To be filled in in a later version.

Acknowledgments

Thanks to Joe Regan and Alfred Nothaft for their contributions.

Authors' Addresses

Yeong-il,Seo
Korea Telecom
Telecommunication Network Laboratory
Member of Technical Staff
463-1 Junmin-dong, Yusung-gu, Taejeon, Korea
Tel : +82-42-870-8333 / FAX : +82-42-870-8339
Mobile : 016-235-0135 / E-mail : syi@hana.ne.kr

Hamid Ould-Brahim
Nortel Networks
P O Box 3511 Station C
Ottawa ON K1Y 4H7 Canada
Phone: +1 (613) 765 3418
Email: hbrahim@nortelnetworks.com

Ashwin Moranganti
Appian Communications
email: amoranganti@appiancom.com
phone: 978 206-7248

Pascal Menezes
Terabeam Networks, Inc.
14833 NE 87th St.
Redmond, WA, USA
phone: (206) 686-2001
email: Pascal.Menezes@Terabeam.com

Pierre Lin
Yipes Communications, Inc.
114 Sansome St.
San Francisco CA 94104

---

email: pierre.lin@yipes.com

Marc Lasserre
Riverstone Networks

5200 Great America Parkway
Santa Clara CA 95054
marc@riverstonenet.com

Sunil Khandekar
TiMetra Networks
274 Ferguson Dr.
Mountain View, CA 94043

Giles Heron
PacketExchange Ltd.
The Truman Brewery
91 Brick Lane
LONDON E1 6QL
United Kingdom
Email: giles@packetexchange.net

Loa Andersson
Utfors AB
Box 525, 169 29 Solna
Sweden
phone: +46 8 5270 5038
email: loa.andersson@utfors.se

Javier Achirica
Telefonica Data
javier.achirica@telefonica-data.com

Vach Kompella
TiMetra Networks
274 Ferguson Dr.
Mountain View, CA 94043
Email: vkompella@timetra.com

Yakov Rekhter
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
yakov@juniper.net

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089

kireeti@juniper.net

IPR Notice

Full Copyright Notice

Acknowledgement

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