Network Working Group
Request for Comments: 2983
Category: Informational

D. Black
EMC Corporation

October 2000

Differentiated Services and Tunnels

Status of this Memo

Copyright Notice

Abstract

This document considers the interaction of Differentiated Services (diffserv) (RFC 2474, RFC 2475) with IP tunnels of various forms. The discussion of tunnels in the diffserv architecture (RFC 2475) provides insufficient guidance to tunnel designers and implementers. This document describes two conceptual models for the interaction of diffserv with Internet Protocol (IP) tunnels and employs them to explore the resulting configurations and combinations of functionality. An important consideration is how and where it is appropriate to perform diffserv traffic conditioning in the presence of tunnel encapsulation and decapsulation. A few simple mechanisms are also proposed that limit the complexity that tunnels would otherwise add to the diffserv traffic conditioning model. Security considerations for IPSec tunnels limit the possible functionality in some circumstances.

1. Conventions used in this document

An IP tunnel encapsulates IP traffic in another IP header as it passes through the tunnel; the presence of these two IP headers is a defining characteristic of IP tunnels, although there may be additional headers inserted between the two IP headers. The inner IP header is that of the original traffic; an outer IP header is attached and detached at tunnel endpoints. In general, intermediate network nodes between tunnel endpoints operate solely on the outer IP header, and hence diffserv-capable intermediate nodes access and modify only the DSCP field in the outer IP header. The terms "tunnel" and "IP tunnel" are used interchangeably in this document. For simplicity, this document does not consider tunnels other than IP tunnels (i.e., for which there is no encapsulating IP header), such

as MPLS paths and "tunnels" formed by encapsulation in layer 2 (link) headers, although the conceptual models and approach described here may be useful in understanding the interaction of diffserv with such tunnels.

This analysis considers tunnels to be unidirectional; bi-directional tunnels are considered to be composed of two unidirectional tunnels carrying traffic in opposite directions between the same tunnel endpoints. A tunnel consists of an ingress where traffic enters the tunnel and is encapsulated by the addition of the outer IP header, an egress where traffic exits the tunnel and is decapsulated by the removal of the outer IP header, and intermediate nodes through which tunneled traffic passes between the ingress and egress. This document does not make any assumptions about routing and forwarding of tunnel traffic, and in particular assumes neither the presence nor the absence of route pinning in any form.

2. Diffserv and Tunnels Overview

Tunnels range in complexity from simple IP-in-IP tunnels [RFC 2003] to more complex multi-protocol tunnels, such as IP in PPP in L2TP in IPSec transport mode [RFC 1661, RFC 2401, RFC 2661]. The most general tunnel configuration is one in which the tunnel is not end-to-end, i.e., the ingress and egress nodes are not the source and destination nodes for traffic carried by the tunnel; such a tunnel may carry traffic with multiple sources and destinations. If the ingress node is the end-to-end source of all traffic in the tunnel, the result is a simplified configuration to which much of the analysis and guidance in this document are applicable, and likewise if the egress node is the end-to-end destination.

A primary concern for differentiated services is the use of the Differentiated Services Code Point (DSCP) in the IP header [RFC 2474, RFC 2475]. The diffserv architecture permits intermediate nodes to examine and change the value of the DSCP, which may result in the DSCP value in the outer IP header being modified between tunnel ingress and egress. When a tunnel is not end-to-end, there are circumstances in which it may be desirable to propagate the DSCP and/or some of the information that it contains to the outer IP header on ingress and/or back to inner IP header on egress. The current situation facing tunnel implementers is that [RFC 2475] offers incomplete guidance. Guideline G.7 in Section 3 is an example, as some PHB specifications have followed it by explicitly specifying the PHBs that may be used in the outer IP header for tunneled traffic. This is overly restrictive; for example, if a specification requires that the same PHB be used in both the inner and outer IP headers, traffic conforming to that specification cannot be tunneled across domains or networks that do not support that PHB.

A more flexible approach that should be used instead is to describe the behavioral properties of a PHB that are important to preserve when traffic is tunneled and allow the outer IP header to be marked in any fashion that is sufficient to preserve those properties.

This document proposes an approach in which traffic conditioning is performed in series with tunnel ingress or egress processing, rather than in parallel. This approach does not create any additional paths that transmit information across a tunnel endpoint, as all diffserv information is contained in the DSCPs in the IP headers. The IPSec architecture [RFC 2401] requires that this be the case to preserve security properties at the egress of IPSec tunnels, but this approach also avoids complicating diffserv traffic conditioning blocks by introducing out-of-band inputs. A consequence of this approach is that the last sentence of Guideline G.7 in Section 3 of [RFC 2475] becomes moot because there are no tunnel egress diffserv components that have access to both the inner and outer DSCPs.

An additional advantage of this traffic conditioning approach is that it places no additional restrictions on the positioning of diffserv domain boundaries with respect to traffic conditioning and tunnel encapsulation/decapsulation components. An interesting class of configurations involves a diffserv domain boundary that passes through (i.e., divides) a network node; such a boundary can be split to create a DMZ-like region between the domains that contains the tunnel encapsulation or decapsulation processing. Diffserv traffic conditioning is not appropriate for such a DMZ-like region, as traffic conditioning is part of the operation and management of diffserv domains.

3. Conceptual Models for Diffserv Tunnels

This analysis introduces two conceptual traffic conditioning models for IP tunnels based on an initial discussion that assumes a fully diffserv-capable network. Configurations in which this is not the case are taken up in Section 3.2.

3.1 Conceptual Models for Fully DS-capable Configurations

The first conceptual model is a uniform model that views IP tunnels as artifacts of the end to end path from a traffic conditioning standpoint; tunnels may be necessary mechanisms to get traffic to its destination(s), but have no significant impact on traffic conditioning. In this model, any packet has exactly one DS Field that is used for traffic conditioning at any point, namely the DS Field in the outermost IP header; any others are ignored. Implementations of this model copy the DSCP value to the outer IP header at encapsulation and copy the outer header's DSCP value to the

inner IP header at decapsulation. Use of this model allows IP tunnels to be configured without regard to diffserv domain boundaries because diffserv traffic conditioning functionality is not impacted by the presence of IP tunnels.

The second conceptual model is a pipe model that views an IP tunnel as hiding the nodes between its ingress and egress so that they do not participate fully in traffic conditioning. In this model, a tunnel egress node uses traffic conditioning information conveyed from the tunnel ingress by the DSCP value in the inner header, and ignores (i.e., discards) the DSCP value in the outer header. The pipe model cannot completely hide traffic conditioning within the tunnel, as the effects of dropping and shaping at intermediate tunnel nodes may be visible at the tunnel egress and beyond.

The pipe model has traffic conditioning consequences when the ingress and egress nodes are in different diffserv domains. In such a situation, the egress node must perform traffic conditioning to ensure that the traffic exiting the tunnel has DSCP values acceptable to the egress diffserv domain (see Section 6 of the diffserv architecture [RFC 2475]). An inter-domain TCA (Traffic Conditioning Agreement) between the diffserv domains containing the tunnel ingress and egress nodes may be used to reduce or eliminate egress traffic conditioning. Complete elimination of egress traffic conditioning requires that the diffserv domains at ingress and egress have compatible service provisioning policies for the tunneled traffic and support all of the PHB groups and DSCP values used for that traffic in a consistent fashion. Examples of this situation are provided by some virtual private network tunnels; it may be useful to view such tunnels as linking the diffserv domains at their endpoints into a diffserv region by making the tunnel endpoints virtually contiguous even though they may be physically separated by intermediate network nodes.

The pipe model is also appropriate for situations in which the DSCP itself carries information through the tunnel. For example, if transit between two domains is obtained via a path that uses the EF PHB [RFC 2598], the drop precedence information in the AF PHB DSCP values [RFC 2597] will be lost unless something is done to preserve it; an IP tunnel is one possible preservation mechanism. A path that crosses one or more non-diffserv domains between its DS-capable endpoints may experience a similar information loss phenomenon if a tunnel is not used due to the limited set of DSCP codepoints that are compatible with such domains.

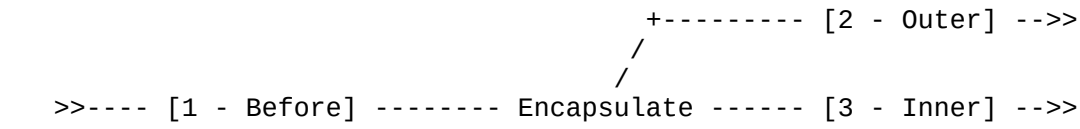
3.2 Considerations for Partially DS-capable Configurations

If only the tunnel egress node is DS-capable, [RFC 2475] requires the egress node to perform any edge traffic conditioning needed by the diffserv domain for tunneled traffic entering from outside the domain. If the egress node would not otherwise be a DS edge node, one way to meet this requirement is to perform edge traffic conditioning at an appropriate upstream DS edge node within the tunnel, and copy the DSCP value from the outer IP header to the inner IP header as part of tunnel decapsulation processing; this applies the uniform model to the portion of the tunnel within the egress node's diffserv domain. A second alternative is to discard the outer DSCP value as part of decapsulation processing, reducing the resulting traffic conditioning problem and requirements to those of an ordinary DS ingress node. This applies the pipe model to the portion of the tunnel within the egress node's diffserv domain and hence the adjacent upstream node for DSCP marking purposes is the tunnel ingress node, rather than the immediately upstream intermediate tunnel node.

If only the tunnel ingress node is DS-capable, [RFC 2475] requires that traffic emerging from the tunnel be compatible with the network at the tunnel egress. If tunnel decapsulation processing discards the outer header's DSCP value without changing the inner header's DSCP value, the DS-capable tunnel ingress node is obligated to set the inner header's DSCP to a value compatible with the network at the tunnel egress. The value 0 (DSCP of 000000) is used for this purpose by a number of existing tunnel implementations. If the egress network implements IP precedence as specified in [RFC 791], then some or all of the eight class selector DSCP codepoints defined in [RFC 2474] may be usable. DSCP codepoints other than the class selectors are not generally suitable for this purpose, as correct operation would usually require diffserv functionality at the DS-incapable tunnel egress node.

4. Ingress Functionality

As described in Section 3 above, this analysis is based on an approach in which diffserv functionality and/or out-of-band communication paths are not placed in parallel with tunnel encapsulation processing. This allows three possible locations for traffic conditioning with respect to tunnel encapsulation processing, as shown in the following diagram that depicts the flow of IP headers through tunnel encapsulation:



Traffic conditioning at [1 - Before] is logically separate from the tunnel, as it is not impacted by the presence of tunnel encapsulation, and hence should be allowed by tunnel designs and specifications. Traffic conditioning at [2 - Outer] may interact with tunnel protocols that are sensitive to packet reordering; such tunnels may need to limit the functionality at [2 - Outer] as discussed further in Section 5.1. In the absence of reordering sensitivity, no additional restrictions should be necessary, although traffic conditioning at [2 - Outer] may be responsible for remarking traffic to be compatible with the next diffserv domain that the tunneled traffic enters.

In contrast, the [3 - Inner] location is difficult to utilize for traffic conditioning because it requires functionality that reaches inside the packet to operate on the inner IP header. This is impossible for IPSec tunnels and any other tunnels that are encrypted or employ cryptographic integrity checks. Hence traffic conditioning at [3 - Inner] can often only be performed as part of tunnel encapsulation processing, complicating both the encapsulation and traffic conditioning implementations. In many cases, the desired functionality can be achieved via a combination of traffic conditioners in the other two locations, both of which can be specified and implemented independently of tunnel encapsulation.

An exception for which traffic conditioning functionality is necessary at [3 - Inner] occurs when the DS-incapable tunnel egress discards the outer IP header as part of decapsulation processing, and hence the DSCP in the inner IP header must be compatible with the egress network. Setting the inner DSCP to 0 as part of encapsulation addresses most of these cases, and the class selector DCSP codepoint values are also useful for this purpose, as they are valid for networks that support IP precedence [RFC 791].

The following table summarizes the achievable relationships among the before (B), outer (O), and inner (I) DSCP values and the corresponding locations of traffic conditioning logic.

Relationship	Traffic Conditioning Location(s)
B = I = O	No traffic conditioning required
B != I = O	[1 - Before]
B = I != O	[2 - Outer]
B = O != I	Limited support as part of encapsulation: I can be set to 000000 or possibly one of the class selector code points.
B != I != O	Some combination of the above three scenarios.

A combination of [1 - Before] and [2 - Outer] is applicable to many cases covered by the last two lines of the table, and may be preferable to deploying functionality at [3 - Inner]. Traffic conditioning may still be required for purposes such as rate and burst control even if DSCP values are not changed.

4.1 Ingress DSCP Selection and Reordering

It may be necessary or desirable to limit the DS behavior aggregates that utilize an IP tunnel that is sensitive to packet reordering within the tunnel. The diffserv architecture allows packets to be reordered when they belong to behavior aggregates among which reordering is permitted; for example, reordering is allowed among behavior aggregates marked with different Class Selector DSCPs [RFC 2474]. IPSec [RFC 2401] and L2TP [RFC 2661] provide examples of tunnels that are sensitive to packet reordering. If IPSec's anti-replay support is configured, audit events are generated in response to packet reordering that exceeds certain levels, with the audit events indicating potential security issues. L2TP can be configured to restore the ingress ordering of packets at tunnel egress, not only undoing any differentiation based on reordering within the tunnel, but also negatively impacting the traffic (e.g., by increasing latency). The uniform model cannot be completely applied to such tunnels, as arbitrary mixing of traffic from different behavior aggregates can cause these undesirable interactions.

The simplest method of avoiding undesirable interactions of reordering with reordering-sensitive tunnel protocols and features is not to employ the reordering-sensitive protocols or features, but this is often not desirable or even possible. When such protocols or features are used, interactions can be avoided by ensuring that the aggregated flows through the tunnel are marked at [2 - Outer] to constitute a single ordered aggregate (i.e., the PHBs used share an ordering constraint that prevents packets from being reordered). Tunnel protocol specifications should indicate both whether and under what circumstances a tunnel should be restricted to a single ordered aggregate as well as the consequences of deviating from that restriction. For the IPSec and L2TP examples discussed above, the

specifications should restrict each tunnel to a single ordered aggregate when protocol features sensitive to reordering are configured, and may adopt the approach of restricting all tunnels in order to avoid unexpected consequences of changes in protocol features or composition of tunneled traffic. Diffserv implementations should not attempt to look within such tunnels to provide reordering-based differentiation to the encapsulated microflows. If reordering-based differentiation is desired within such tunnels, multiple parallel tunnels between the same endpoints should be used. This enables reordering among packets in different tunnels to coexist with an absence of packet reordering within each individual tunnel. For IPSec and related security protocols, there is no cryptographic advantage to using a single tunnel for multiple ordered aggregates rather than multiple tunnels because any traffic analysis made possible by the use of multiple tunnels can also be performed based on the DSCPs in the outer headers of traffic in a single tunnel. In general, the additional resources required to support multiple tunnels (e.g., cryptographic contexts), and the impact of multiple tunnels on network management should be considered in determining whether and where to deploy them.

4.2 Tunnel Selection

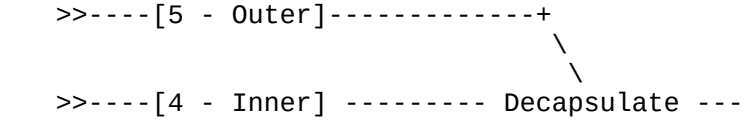
The behavioral characteristics of a tunnel are an important consideration in determining what traffic should utilize the tunnel. This involves the service provisioning policies of all the participating domains, not just the PHBs and DSCPs marked on the traffic at [2 - Outer]. For example, while it is in general a bad idea to tunnel EF PHB traffic via a Default PHB tunnel, this can be acceptable if the EF traffic is the only traffic that utilizes the tunnel, and the tunnel is provisioned in a fashion adequate to preserve the behavioral characteristics required by the EF PHB.

Service provisioning policies are responsible for preventing mismatches such as forwarding EF traffic via an inadequately provisioned Default tunnel. When multiple parallel tunnels with different behavioral characteristics are available, service provisioning policies are responsible for determining which flows should use which tunnels. Among the possibilities is a coarse version of the uniform tunnel model in which the inner DSCP value is used to select a tunnel that will forward the traffic using a behavioral aggregate that is compatible with the traffic’s PHB.

5. Egress Functionality

As described in Section 3 above, this analysis is based on an approach in which diffserv functionality and/or out-of-band communication paths are not placed in parallel with tunnel

encapsulation processing. This allows three possible locations for traffic conditioners with respect to tunnel decapsulation processing, as shown in the following diagram that depicts the flow of IP headers through tunnel decapsulation:



Traffic conditioning at [5 - Outer] and [6 - After] is logically separate from the tunnel, as it is not impacted by the presence of tunnel decapsulation. Tunnel designs and specifications should allow diffserv traffic conditioning at these locations. Such conditioning can be viewed as independent of the tunnel, i.e., [5 - Outer] is traffic conditioning that takes place prior to tunnel egress, and [6 - After] is traffic conditioning that takes place after egress decapsulation. An important exception is that the configuration of a tunnel (e.g., the absence of traffic conditioning at tunnel ingress) and/or the diffserv domains involved may require that all traffic exiting a tunnel pass through diffserv traffic conditioning to fulfill the diffserv edge node traffic conditioning responsibilities of the tunnel egress node. Tunnel designers are strongly encouraged to include the ability to require that all traffic exiting a tunnel pass through diffserv traffic conditioning in order to ensure that traffic exiting the node is compatible with the egress node's diffserv domain.

In contrast, the [4 - Inner] location is difficult to employ for traffic conditioning because it requires reaching inside the packet to operate on the inner IP header. Unlike the [3 - Inner] case for encapsulation, there is no need for functionality to be performed at [4- Inner], as diffserv traffic conditioning can be appended to the tunnel decapsulation (i.e., performed at [6 - After]).

5.1 Egress DSCP Selection

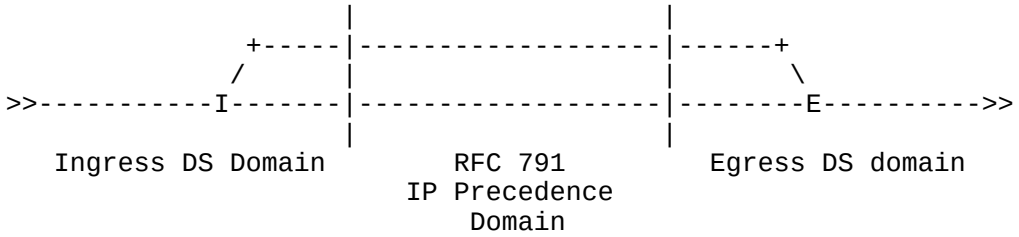
The elimination of parallel functionality and data paths from decapsulation causes a potential loss of information. As shown in the above diagram, decapsulation combines and reduces two DSCP values to one DSCP value, losing information in the most general case, even if arbitrary functionality is allowed. Beyond this, allowing arbitrary functionality poses a structural problem, namely that the DSCP value from the outer IP header would have to be presented as an out-of-band input to the traffic conditioning block at [6 - After], complicating the traffic conditioning model.

To avoid such complications, the simpler approach of statically selecting either the inner or outer DSCP value at decapsulation is recommended, leaving the full generality of traffic conditioning functionality to be implemented at [5 - Outer] and/or [6 - After]. Tunnels should support static selection of one or the other DSCP value at tunnel egress. The rationale for this approach is usually only one of the two DSCP values contains useful information. The conceptual model for the tunnel provides a strong indication of which one contains useful information; the outer DSCP value usually contains the useful information for tunnels based on the uniform model, and the inner DSCP value usually contains the useful information for tunnels based on the pipe model. IPSec tunnels are usually based on the pipe model, and for security reasons are currently required to select the inner DSCP value; they should not be configured to select the outer DSCP value in the absence of an adequate security analysis of the resulting risks and implications.

5.2 Egress DSCP Selection Case Study

As a sanity check on the egress DSCP selection approach proposed above, this subsection considers a situation in which a more complex approach might be required. Statically choosing a single DSCP value may not work well when both DSCPs are carrying information that is relevant to traffic conditioning.

As an example, consider a situation in which different AF groups [RFC 2597] are used by the two domains at the tunnel endpoints, and there is an intermediate domain along the tunnel using RFC 791 IP precedences that is transited by setting the DSCP to zero. This situation is shown in the following IP header flow diagram where I is the tunnel ingress node, E is the tunnel egress node and the vertical lines are domain boundaries. The node at the left-hand vertical line sets the DSCP in the outer header to 0 in order to obtain compatibility with the middle domain:



In this situation, the DS edge node for the egress domain (i.e., the node at the right-hand vertical line) can select the appropriate AF group (e.g., via an MF classifier), but cannot reconstruct the drop precedence information that was removed from the outer header when it

transited the RFC 791 domain (although it can construct new information via metering and marking). The original drop precedence information is preserved in the inner IP header's DSCP, and could be combined at the tunnel egress with the AF class selection communicated via the outer IP header's DSCP. The marginal benefit of being able to reuse the original drop precedence information as opposed to constructing new drop precedence markings does not justify the additional complexity introduced into tunnel egress traffic conditioners by making both DSCP values available to traffic conditioning at [6 - After].

6. Diffserv and Protocol Translators

A related issue involves protocol translators, including those employing the Stateless IP/ICMP Translation Algorithm [RFC 2765]. These translators are not tunnels because they do not add or remove a second IP header to/from packets (e.g., in contrast to IPv6 over IPv4 tunnels [RFC 1933]) and hence do not raise concerns of information propagation between inner and outer IP headers. The primary interaction between translators and diffserv is that the translation boundary is likely to also be a diffserv domain boundary (e.g., the IPv4 and IPv6 domains may have different policies for traffic conditioning and DSCP usage), and hence such translators should allow the insertion of diffserv edge node processing (including traffic conditioning) both before and after the translation processing.

7. Security Considerations

The security considerations for the diffserv architecture discussed in [RFC 2474, RFC 2475] apply when tunnels are present. One of the requirements is that a tunnel egress node in the interior of a diffserv domain is the DS ingress node for traffic exiting the tunnel, and is responsible for performing appropriate traffic conditioning. The primary security implication is that the traffic conditioning is responsible for dealing with theft- and denial-of-service threats posed to the diffserv domain by traffic exiting from the tunnel. The IPSec architecture [RFC 2401] places a further restriction on tunnel egress processing; the outer header is to be discarded unless the properties of the traffic conditioning to be applied are known and have been adequately analyzed for security vulnerabilities. This includes both the [5 - Outer] and [6 - After] traffic conditioning blocks on the tunnel egress node, if present, and may involve traffic conditioning performed by an upstream DS-edge node that is the DS domain ingress node for the encapsulated tunneled traffic.

8. References

[RFC 791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[RFC 1661] Simpson, W., "The Point-to-Point Protocol (PPP)", STD 51, RFC 1661, July 1994.

[RFC 1933] Gilligan, R. and E. Nordmark, "Transition Mechanisms for IPv6 Hosts and Routers", RFC 1933, April 1996.

[RFC 2003] Perkins, C., "IP Encapsulation within IP", RFC 2003, October 1996.

[RFC 2401] Kent, S. and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

[RFC 2474] Nichols, K., Blake, S., Baker, F. and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC 2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z. and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC 2597] Heinanen, J., Baker, F., Weiss, W. and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597. June 1999.

[RFC 2598] Jacobson, V., Nichols, K. and K. Poduri, "An Expedited Forwarding PHB", RFC 2598, June 1999.

[RFC 2661] Townsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G. and B. Palter. "Layer Two Tunneling Protocol "L2TP"", RFC 2661, August 1999.

[RFC 2765] Nordmark, E., "Stateless IP/ICMP Translation Algorithm (SIIT)", RFC 2765, February 2000.

9. Acknowledgments

Some of this material is based on discussions with Brian Carpenter, and in particular his presentation on this topic to the diffserv WG during the summer 1999 IETF meeting in Oslo. Credit is also due to a number of people working on tunnel specifications who have discovered limitations of the diffserv architecture [RFC 2475] in the area of tunnels. Their patience with the time it has taken to address this set of issues is greatly appreciated. Finally, this material has benefited from discussions within the diffserv WG, both in meetings and on the mailing list -- the contributions of participants in those discussions are gratefully acknowledged.

10. Author's Address

David L. Black
EMC Corporation
42 South St.
Hopkinton, MA 01748

Phone: +1 (508) 435-1000 x75140
EMail: black_david@emc.com

11. Full Copyright Statement

Acknowledgement

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