Y. Bernet, Microsoft
R. Yavatkar, Intel
P. Ford, Microsoft
F. Baker, Cisco
L. Zhang, UCLA
M. Speer, Sun Microsystems
R. Braden, ISI
B. Davie, Cisco
John Wroclawski, MIT LCS
E. Felstaine, Allot Communications

Internet Draft
Expires: September, 2000
Document: draft-ietf-issll-diffserv-rsvp-04.txt March, 2000

A Framework For Integrated Services Operation Over Diffserv Networks

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026. Internet-Drafts are Working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet- Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Copyright Notice

Copyright (C) The Internet Society (2000). All Rights Reserved.

1. Abstract

The Integrated Services architecture provides a means for the delivery of end-to-end QoS to applications over heterogeneous networks. To support this end-to-end model, the Intserv architecture must be supported over a wide variety of different types of network elements. In this context, a network that supports Differentiated Services (Diffserv) may be viewed as a network element in the total end-to-end path. This document describes a framework by which Integrated Services may be supported over Diffserv networks.

2. Introduction

Work on QoS-enabled IP networks has led to two distinct approaches: the Integrated Services architecture (Intserv)[10] and its accompanying signaling protocol, RSVP [1], and the Differentiated Services architecture (Diffserv)[8]. This document describes ways in which a Diffserv network can be used in the context of the Intserv architecture to support the delivery of end-to-end QOS.

2.1 Integrated Services Architecture

The integrated services architecture defined a set of extensions to the traditional best effort model of the Internet with the goal of allowing end-to-end QOS to be provided to applications. One of the key components of the architecture is a set of service definitions; the current set of services consists of the controlled load and guaranteed services. The architecture assumes that some explicit setup mechanism is used to convey information to routers so that they can provide requested services to flows that require them. While RSVP is the most widely known example of such a setup mechanism, the Intserv architecture is designed to accommodate other mechanisms.

Intserv services are implemented by "network elements". While it is common for network elements to be individual nodes such as routers or links, more complex entities, such as ATM "clouds" or 802.3 networks may also function as network elements. As discussed in more detail below, a Diffserv network (or "cloud") may be viewed as a network element within a larger Intserv network.

2.3 RSVP

RSVP is a signaling protocol that applications may use to request resources from the network. The network responds by explicitly admitting or rejecting RSVP requests. Certain applications that have quantifiable resource requirements express these requirements using Intserv parameters as defined in the appropriate Intserv service specification. As noted above, RSVP and Intserv are separable. RSVP is a signaling protocol which may carry Intserv information. Intserv defines the models for expressing service types, quantifying resource requirements and for determining the availability of the requested resources at relevant network elements (admission control).

The current prevailing model of RSVP usage is based on a combined

RSVP/Intserv architecture. In this model, RSVP signals per-flow resource requirements to network elements, using Intserv parameters. These network elements apply Intserv admission control to signaled requests. In addition, traffic control mechanisms on the network element are configured to ensure that each admitted flow receives the service requested in strict isolation from other traffic. To

this end, RSVP signaling configures microflow (MF) [8] packet classifiers in Intserv capable routers along the path of the traffic flow. These classifiers enable per-flow classification of packets based on IP addresses and port numbers.

The following factors have impeded deployment of RSVP (and the Intserv architecture) in the Internet at large:

1. The use of per-flow state and per-flow processing raises scalability concerns for large networks.

2. Only a small number of hosts currently generate RSVP signaling. While this number is expected to grow dramatically, many applications may never generate RSVP signaling.

3. The necessary policy control mechanisms -- access control, authentication, and accounting -- have only recently become available [17].

2.4 Diffserv

In contrast to the per-flow orientation of RSVP, Diffserv networks classify packets into one of a small number of aggregated flows or "classes", based on the Diffserv codepoint (DSCP) in the packet's IP header. This is known as behavior aggregate (BA) classification [8]. At each Diffserv router, packets are subjected to a "per-hop behavior" (PHB), which is invoked by the DSCP. The primary benefit of Diffserv is its scalability. Diffserv eliminates the need for per-flow state and per-flow processing and therefore scales well to large networks.

2.5 Roles of Intserv, RSVP and Diffserv

We view Intserv, RSVP and Diffserv as complementary technologies in the pursuit of end-to-end QoS. Together, these mechanisms can facilitate deployment of applications such as IP-telephony, video-on-demand, and various non-multimedia mission-critical applications. Intserv enables hosts to request per-flow, quantifiable resources, along end-to-end data paths and to obtain feedback regarding admissibility of these requests. Diffserv enables scalability across

large networks.

2.6 Components of Intserv, RSVP and Diffserv

Before proceeding, it is helpful to identify the following components of the QoS technologies described:

RSVP signaling - This term refers to the standard RSVP signaling protocol. RSVP signaling is used by hosts to signal application resource requirements to the network (and to each other). Network elements use RSVP signaling to return an admission control decision to hosts. RSVP signaling may or may not carry Intserv parameters.

Admission control at a network element may or may not be based on the Intserv model.

MF traffic control - This term refers to traffic control which is applied independently to individual traffic flows and therefore requires recognizing individual traffic flows via MF classification.

Aggregate traffic control - This term refers to traffic control which is applied collectively to sets of traffic flows. These sets of traffic flows are recognized based on BA (DSCP) classification. In this document, we use the terms "aggregate traffic control" and "Diffserv" interchangeably.

Aggregate RSVP. While the existing definition of RSVP supports only per-flow reservations, extensions to RSVP are being developed to enable RSVP reservations to be made for aggregated traffic, i.e. sets of flows that may be recognized by BA classification. This use of RSVP may be useful in controlling the allocation of bandwidth in Diffserv networks.

Per-flow RSVP. The conventional usage of RSVP to perform resource reservations for individual microflows.

RSVP/Intserv - This term is used to refer to the prevailing model of RSVP usage which includes RSVP signaling with Intserv parameters, Intserv admission control and per-flow traffic control at network elements.

Diffserv Region. A set of contiguous routers which support BA classification and traffic control. While such a region may also support MF classification, the goal of this document is to describe how such a region may be used in delivery of end-to-end QOS when only BA classification is performed inside the Diffserv region.

Non-Diffserv Region. The portions of the network outside the Diffserv region. Such a region may also offer a variety of different types of classification and traffic control.

Note that, for the purposes of this document, the defining features of a Diffserv region is the type of classification and traffic control that is used for the delivery of end-to-end QOS for a particular application. Thus, while it may not be possible to identify a certain region as "purely Diffserv" with respect to all traffic flowing through the region, it is possible to define it in this way from the perspective of the treatment of traffic from a single application.

2.7 The Framework

In the framework we present, end-to-end, quantitative QoS is provided by applying the Intserv model end-to-end across a network

containing one or more Diffserv regions. The Diffserv regions may, but are not required to, participate in end-to-end RSVP signaling for the purpose of optimizing resource allocation and supporting admission control.

From the perspective of Intserv, Diffserv regions of the network are treated as virtual links connecting Intserv capable routers or hosts (much as an 802.1p network region is treated as a virtual link in [5]). Within the Diffserv regions of the network routers implement specific PHBs (aggregate traffic control). The total amount of traffic that is admitted into the Diffserv region that will receive a certain PHB may be limited by policing at the edge. As a result we expect that the Diffserv regions of the network will be able to support the Intserv style services requested from the periphery. In our framework, we address the support of end-to-end Integrated Services over the Diffserv regions of the network. Our goal is to enable seamless inter-operation. As a result, the network administrator is free to choose which regions of the network act as Diffserv regions. In one extreme the Diffserv region is pushed all the way to the periphery, with hosts alone having full Intserv capability. In the other extreme, Intserv is pushed all the way to the core, with no Diffserv region.

2.8 Contents

In section 3 we discuss the benefits that can be realized by using the aggregate traffic control provided by Diffserv network regions in the broader context of the Intserv architecture. In section 4, we

present the framework and the reference network. Section 5 details two possible realizations of the framework. Section 6 discusses the implications of the framework for Diffserv. Section 7 presents some issues specific to multicast flows.

3. Benefits of Using Intserv with Diffserv

The primary benefit of Diffserv aggregate traffic control is its scalability. In this section, we discuss the benefits that interoperation with Intserv can bring to a Diffserv network region. Note that this discussion is in the context of servicing quantitative QoS applications specifically. By this we mean those applications that are able to quantify their traffic and QoS requirements.

3.1 Resource Based Admission Control

In Intserv networks, quantitative QoS applications use an explicit setup mechanism (e.g. RSVP) to request resources from the network. The network may accept or reject these requests in response. This is "explicit admission control". Explicit admission control helps to assure that network resources are optimally used. To further understand this issue, consider a Diffserv network region providing only aggregate traffic control with no signaling. In the Diffserv

network region, admission control is applied implicitly by provisioning policing parameters at network elements. For example, a network element at the ingress to a Diffserv network region could be provisioned to accept only 50 Kbps of traffic for the EF DSCP.

While such implicit admission control does protect the network to some degree, it can be quite ineffective. For example, consider that there may be 10 IP telephony sessions originating outside the Diffserv network region, each requiring 10 Kbps of EF service from the Diffserv network region. Since the network element protecting the Diffserv network region is provisioned to accept only 50 Kbps of traffic for the EF DSCP, it will discard half the offered traffic. This traffic will be discarded from the aggregation of traffic marked EF, with no regard to the microflow from which it originated. As a result, it is likely that of the ten IP telephony sessions, none will obtain satisfactory service when in fact, there are sufficient resources available in the Diffserv network region to satisfy five sessions.

In the case of explicit admission control, the network will signal rejection in response to requests for resources that would exceed the 50 Kbps limit. As a result, upstream network elements (including

originating hosts) and applications will have the information they require to take corrective action. The application might respond by refraining from transmitting, or by requesting admission for a lesser traffic profile. The host operating system might respond by marking the application's traffic for the DSCP that corresponds to best-effort service. Upstream network elements might respond by re-marking packets on the rejected flow to a lower service level. In some cases, it may be possible to reroute traffic over alternate paths or even alternate networks (e.g. the PSTN for voice calls). In any case, the integrity of those flows that were admitted would be preserved, at the expense of the flows that were not admitted. Thus, by appointing an Intserv-conversant admission control agent for the Diffserv region of the network it is possible to enhance the service that the network can provide to quantitative QoS applications.

3.2 Policy Based Admission Control

In network regions where RSVP is used, resource requests can be intercepted by RSVP-aware network elements and can be reviewed against policies stored in policy databases. These resource requests securely identify the user and the application for which the resources are requested. Consequently, the network element is able to consider per-user and/or per-application policy when deciding whether or not to admit a resource request. So, in addition to optimizing the use of resources in a Diffserv network region (as discussed in 3.1) RSVP conversant admission control agents can be used to apply specific customer policies in determining the specific customer traffic flows entitled to use the Diffserv network region's resources. Customer policies can be used to allocate resources to specific users and/or applications.

By comparison, in Diffserv network regions without RSVP signaling, policies are typically applied based on the Diffserv customer network from which traffic originates, not on the originating user or application within the customer network.

3.3 Assistance in Traffic Identification/Classification

Within Diffserv network regions, traffic is allotted service based on the DSCP marked in each packet's IP header. Thus, in order to obtain a particular level of service within the Diffserv network region, it is necessary to effect the marking of the correct DSCP in packet headers. There are two mechanisms for doing so, host marking and router marking. In the case of host marking, the host operating system marks the DSCP in transmitted packets. In the case of router

marking, routers in the network are configured to identify specific traffic (typically based on MF classification) and to mark the DSCP as packets transit the router. There are advantages and disadvantages to each scheme. Regardless of the scheme used, explicit signaling offers significant benefits.

3.3.1 Host Marking

In the case of host marking, the host operating system marks the DSCP in transmitted packets. This approach has the benefit of shifting per-flow classification and marking to the source of the traffic, where it scales best. It also enables the host to make decisions regarding the mark that is appropriate for each transmitted packet and hence the relative importance attached to each packet. The host is generally better equipped to make this decision than the network. Furthermore, if IPSEC encryption is used, the host may be the only device in the network that is able to make a meaningful determination of the appropriate marking for each packet.

Host marking requires that the host be aware of the interpretation of DSCPs by the network. This information can be configured into each host. However, such configuration imposes a management burden. Alternatively, hosts can use an explicit signaling protocol such as RSVP to query the network to obtain a suitable DSCP or set of DSCPs to apply to packets for which a certain Intserv service has been requested. An example of how this can be achieved is described in [14].

3.3.2 Router Marking

In the case of router marking, MF classification criteria must be configured in the router. This may be done dynamically, by request from the host operating system, or statically via manual configuration or via automated scripts.

There are significant difficulties in doing so statically. Typically, it is desirable to allot service to traffic based on the application and/or user originating the traffic. At times it is possible to identify packets associated with a specific application by the IP port numbers in the headers. It may also be possible to identify packets originating from a specific user by the source IP address. However, such classification criteria may change frequently. Users may be assigned different IP addresses by DHCP. Applications may use transient ports. To further complicate matters,

multiple users may share an IP address. These factors make it very difficult to manage static configuration of the classification information required to mark traffic in routers.

An attractive alternative to static configuration is to allow host operating systems to signal classification criteria to the router on behalf of users and applications. As we will show later in this document, RSVP signaling is ideally suited for this task. In addition to enabling dynamic and accurate updating of MF classification criteria, RSVP signaling enables classification of IPSEC [13] packets (by use of the SPI) which would otherwise be unrecognizable.

3.4 Traffic Conditioning

Intserv-capable network elements are able to condition traffic at a per-flow granularity, by some combination of shaping and/or policing. Pre-conditioning traffic in this manner before it is submitted to the Diffserv region of the network is beneficial. In particular, it enhances the ability of the Diffserv region of the network to provide quantitative services using aggregate traffic control.

4. The Framework

In the general framework we envision an Internet in which the Integrated Services architecture is used to deliver end-to-end QOS to applications. The network includes some combination of Intserv capable nodes (in which MF classification and per-flow traffic control is applied) and Diffserv regions (in which aggregate traffic control is applied). Individual routers may or may not participate in RSVP signaling regardless of where in the network they reside.

We will consider two specific realizations of the framework. In the first, resources within the Diffserv regions of the network are statically provisioned and these regions include no RSVP aware devices. In the second, resources within the Diffserv region of the network are dynamically provisioned and select devices within the Diffserv network regions participate in RSVP signaling.

4.1 Reference Network

The two realizations of the framework will be discussed in the context of the following reference network:

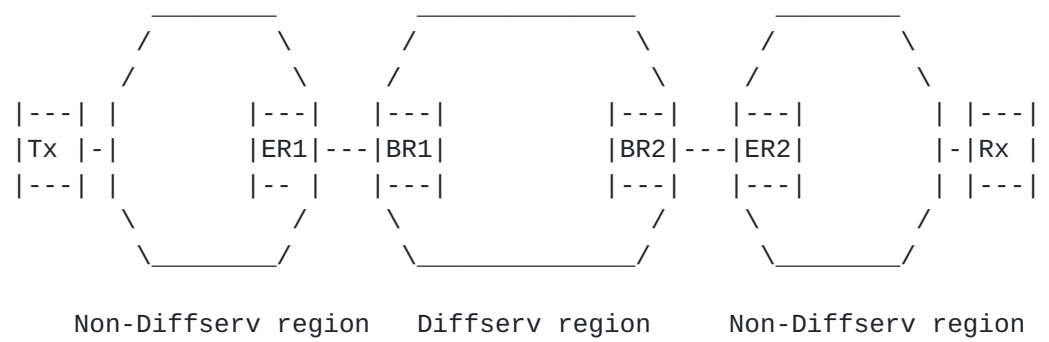


Figure 1: Sample Network Configuration

The reference network includes a Diffserv region in the middle of a larger network supporting Intserv end-to-end. The Diffserv region contains a mesh of routers, at least some of which provide aggregate traffic control. The regions outside the Diffserv region (non-Diffserv regions) contain meshes of routers and attached hosts, at least some of which support the Integrated Services architecture.

In the interest of simplicity we consider a single QoS sender, Tx communicating across this network with a single QoS receiver, Rx. The edge routers (ER1, ER2) which are adjacent to the Diffserv region interface to the border routers (BR1, BR2) within the Diffserv region.

From an economic viewpoint, we may consider that the Diffserv region sells service to the network outside the Diffserv region, which in turn provides service to hosts. Thus, we may think of the non-Diffserv regions as clients or customers of the Diffserv region. In the following, we use the term "customer" for the non-Diffserv regions. Note that the boundaries of the regions may or may not align with administrative domain boundaries, and that a single region might contain multiple administrative domains.

We now define the major components of the reference network.

4.1.1 Hosts

We assume that both sending and receiving hosts use RSVP to communicate the quantitative QoS requirements of QoS-aware applications running on the host. In principle, other mechanisms may be used to establish resource reservations in Intserv-capable nodes, but RSVP is clearly the prevalent mechanism for this purpose.

Typically, a QoS process within the host operating system generates RSVP signaling on behalf of applications. This process may also invoke local traffic control.

As discussed above, traffic control in the host may mark the DSCP in transmitted packets, and shape transmitted traffic to the requirements of the Intserv service in use. Alternatively, the first Intserv-capable router downstream from the host may provide these traffic control functions.

4.1.2 End-to-End RSVP Signaling

We assume that RSVP signaling messages travel end-to-end between hosts Tx and Rx to support RSVP/Intserv reservations outside the Diffserv network region. We require that these end-to-end RSVP messages are at least carried across the Diffserv region. Depending on the specific realization of the framework, these messages may be processed by none, some or all of the routers in the Diffserv region.

4.1.3 Edge Routers

ER1 and ER2 are edge routers, residing adjacent to the Diffserv network regions. The functionality of the edge routers varies depending on the specific realization of the framework. In the case in which the Diffserv network region is RSVP unaware, edge routers act as admission control agents to the Diffserv network. They process signaling messages from both Tx and Rx, and apply admission control based on resource availability within the Diffserv network region and on customer defined policy. In the case in which the Diffserv network region is RSVP aware, the edge routers apply admission control based on local resource availability and on customer defined policy. In this case, the border routers act as the admission control agent to the Diffserv network region.

We will later describe the functionality of the edge routers in greater depth for each of the two realizations of the framework.

4.1.4 Border Routers

BR1 and BR2 are border routers, residing in the Diffserv network region. The functionality of the border routers varies depending on the specific realization of the framework. In the case in which the Diffserv network region is RSVP-unaware, these routers act as pure Diffserv routers. As such, their sole responsibility is to police submitted traffic based on the service level specified in the DSCP and the agreement negotiated with the customer (aggregate traffic control). In the case in which the Diffserv network region is RSVP-aware, the border routers participate in RSVP signaling and act as admission control agents for the Diffserv network region.

We will later describe the functionality of the border routers in greater depth for each of the two realizations of the framework.

4.1.5 Diffserv Network Region

The Diffserv network region supports aggregate traffic control and is assumed not to be capable of MF classification. Depending on the specific realization of the framework, some number of routers within the Diffserv region may be RSVP aware and therefore capable of per-flow signaling and admission control. If devices in the Diffserv region are not RSVP aware, they will pass RSVP messages transparently with negligible performance impact (see [6]).

The Diffserv network region provides two or more levels of service based on the DSCP in packet headers. It may be a single administrative domain or may span multiple domains.

4.1.5 Non-Diffserv Network Regions

The network outside of the Diffserv region consists of Intserv capable hosts and other network elements. Other elements may include routers and perhaps various types of network (e.g. 802, ATM, etc.). These network elements may reasonably be assumed to support Intserv, although this might not be required in the case of over-provisioning. Even if these elements are not Intserv capable, we assume that they will pass RSVP messages unhindered. Routers outside of the Diffserv network region are not precluded from providing aggregate traffic control to some subset of the traffic passing through them.

4.2 Service Mapping

Intserv service requests specify an Intserv service type and a set of quantitative parameters known as a "flowspec". At each hop in an Intserv network, the Intserv service requests are interpreted in a form meaningful to the specific link layer medium. For example at an 802.1 hop, the Intserv parameters are mapped to an appropriate 802.1p priority level [5].

In our framework, Diffserv regions of the network are analogous to the 802.1p capable switched segments described in [5]. Requests for Intserv services must be mapped onto the underlying capabilities of the Diffserv network region. Aspects of the mapping include:

- selecting an appropriate PHB, or set of PHBs, for the requested service;

- performing appropriate policing (including, perhaps, shaping or remarking) at the edges of the Diffserv region;
- exporting Intserv parameters from the Diffserv region (e.g. for the updating of ADSPECs);
- performing admission control on the Intserv requests that takes into account the resource availability in the Diffserv region.

Exactly how these functions are performed will be a function of the way bandwidth is managed inside the Diffserv network region, which is a topic we discuss in Section 4.3.

When the PHB (or set of PHBs) has been selected for a particular Intserv flow, it may be necessary to communicate the choice of DSCP for the flow to other network elements. Two schemes may be used to achieve this end, as discussed below.

4.2.1 Default Mapping

In this scheme, there is some standard, well-known mapping from Intserv service type to a DSCP that will invoke the appropriate behavior in the Diffserv network.

4.2.2 Network Driven Mapping

In this scheme, RSVP conversant routers in the Diffserv network region (perhaps at its edge) may override the well-known mapping described in 4.2.1. In the case that DSCPs are marked at the ingress to the Diffserv region, the DSCPs can simply be remarked at the boundary routers. However, in the case that DSCP marking occurs upstream of the Diffserv region, either in a host or a router, then the appropriate mapping needs to be communicated upstream, to the marking device. This may be accomplished using RSVP, as described in [14].

The decision regarding where to mark DSCP and whether to override the well-known service mapping is a mater of policy to be decided by the administrator of the Diffserv network region in cooperation with the administrator of the network adjacent to the Diffserv region.

4.2.3 Microflow Separation

Boundary routers residing at the edge of the Diffserv region will typically police traffic submitted from the outside the Diffserv region in order to protect resources within the Diffserv region.

This policing will be applied on an aggregate basis, with no regard for the individual microflows making up each aggregate. As a result, it is possible for a misbehaving microflow to claim more than its fair share of resources within the aggregate, thereby degrading the service provided to other microflows. This problem may be addressed by:

1. Providing per microflow policing at the edge routers - this is generally the most appropriate location for microflow policing, since it pushes per-flow work to the edges of the network, where it scales better. In addition, since Intserv-capable routers outside the Diffserv region are responsible for providing microflow service to their customers and the Diffserv region is responsible for providing aggregate service to its customers, this distribution of functionality mirrors the distribution of responsibility.

2. Providing per microflow policing at the border routers - this approach tends to be less scalable than the previous approach. It also imposes a management burden on the Diffserv region of the network. However, it may be appropriate in certain cases, for the Diffserv boundary routers to offer per microflow policing as a value-add to its Intserv customers.

3. Relying on upstream shaping and policing - in certain cases, the customer may trust the shaping of certain groups of hosts sufficiently to not warrant reshaping or policing at the boundary of the Diffserv region. Note that, even if the hosts are shaping microflows properly, these shaped flows may become distorted as they transit through the non-Diffserv region of the network. Depending on the degree of distortion, it may be necessary to somewhat over-provision the aggregate capacities in the Diffserv region, or to re-police using either 1 or 2 above.

The choice of one mechanism or another is a matter of policy to be decided by the administrator of the network outside the Diffserv region.

4.3 Resource Management in Diffserv Regions

A variety of options exist for management of resources (e.g., bandwidth) in the Diffserv network regions to meet the needs of end-to-end Intserv flows. These options include:

- statically provisioned resources;
- resources dynamically provisioned by RSVP;
- resources dynamically provisioned by other means (e.g., a form of Bandwidth Broker).

Some of the details of using each of these different approaches are discussed in the following section.

5. Detailed Examples of the Operation of Intserv over Diffserv Regions

In this section we provide detailed examples of our framework in action. We discuss two examples, one in which the Diffserv network region is RSVP unaware, the other in which the Diffserv network region is RSVP aware.

5.1 Statically Provisioned Diffserv Network Region

In this example, no devices in the Diffserv network region are RSVP aware. The Diffserv network region is statically provisioned. The customer(s) of the Diffserv network regions and the owner of the Diffserv network region have negotiated a static contract (service level specification, or SLS) for the transmit capacity to be provided to the customer at each of a number of standard Diffserv service levels. The "transmit capacity" may be simply an amount of

bandwidth or it could be a more complex "profile" involving a number of factors such as burst size, peak rate, time of day etc.

It is helpful to consider each edge router in the customer network as consisting of two halves, a standard Intserv half, which interfaces to the customer's network regions and a Diffserv half which interfaces to the Diffserv network region. The Intserv half is able to identify and process traffic on per-flow granularity.

The Diffserv half of the router can be considered to consist of a number of virtual transmit interfaces, one for each Diffserv service level negotiated in the SLS. The router contains a table that indicates the transmit capacity provisioned, per the SLS at each Diffserv service level. This table, in conjunction with the default mapping described in 4.2.1, is used to perform admission control decisions on Intserv flows which cross the Diffserv network region.

5.1.1 Sequence of Events in Obtaining End-to-end QoS

The following sequence illustrates the process by which an application obtains end-to-end QoS when RSVP is used by the hosts.

1. The QoS process on the sending host Tx generates an RSVP PATH message that describes the traffic offered by the sending application.

2. The PATH message is carried toward the receiving host, Rx. In the network region to which the sender is attached, standard RSVP/Intserv processing is applied at capable network elements.

3. At the edge router ER1, the PATH message is subjected to standard RSVP processing and PATH state is installed in the router. The PATH message is sent onward to the Diffserv network region.

4. The PATH message is ignored by routers in the Diffserv network region and then processed at ER2 according to standard RSVP processing rules.

5. When the PATH message reaches the receiving host Rx, the operating system generates an RSVP RESV message, indicating interest in offered traffic of a certain Intserv service type.

6. The RESV message is carried back towards the Diffserv network region and the sending host. Consistent with standard RSVP/Intserv processing, it may be rejected at any RSVP-capable node in the path if resources are deemed insufficient to carry the traffic requested.

7. At ER2, the RESV message is subjected to standard RSVP/Intserv processing. It may be rejected if resources on the downstream interface of ER2 are deemed insufficient to carry the resources requested. If it is not rejected, it will be carried transparently through the Diffserv network region, arriving at ER1.

8. In ER1, the RESV message triggers admission control processing. ER1 compares the resources requested in the RSVP/Intserv request to the resources available in the Diffserv network region at the corresponding Diffserv service level. The corresponding service level is determined by the Intserv to Diffserv mapping discussed previously. The availability of resources is determined by the capacity provisioned in the SLS. ER1 may also apply a policy decision such that the resource request may be rejected based on the customer's specific policy criteria, even though the aggregate resources are determined to be available per the SLS.

9. If ER1 approves the request, the RESV message is admitted and is allowed to continue upstream towards the sender. If it rejects the request, the RESV is not forwarded and the appropriate RSVP error messages are sent. If the request is approved, ER1 updates its internal tables to indicate the reduced capacity available at the admitted service level on its transmit interface.

10. The RESV message proceeds through the network region to which the sender is attached. Any RSVP node in this region may reject the reservation request due to inadequate resources or policy. If the request is not rejected, the RESV message will arrive at the sending host, Tx.

11. At Tx, the QoS process receives the RESV message. It interprets receipt of the message as indication that the specified traffic flow has been admitted for the specified Intserv service type (in the Intserv-capable nodes). It may also learn the appropriate DSCP marking to apply to packets for this flow from information provided in the RESV.

12. Tx may mark the DSCP in the headers of packets that are transmitted on the admitted traffic flow. The DSCP may be the default value which maps to the Intserv service type specified in the admitted RESV message, or it may be a value explicitly provided in the RESV.

In this manner, we obtain end-to-end QoS through a combination of networks that support RSVP/Intserv and networks that support Diffserv.

5.2 RSVP-Aware Diffserv Network Region

In this example, the customer's edge routers are standard RSVP routers. The border router, BR1 is RSVP aware. In addition, there may be other routers within the Diffserv network region which are RSVP aware. Note that although these routers are able to participate in some form of RSVP signaling, they classify and schedule traffic in aggregate, based on DSCP, not on the per-flow classification criteria used by standard RSVP/Intserv routers. It can be said that their control-plane is RSVP while their data-plane is Diffserv. This

approach exploits the benefits of RSVP signaling while maintaining much of the scalability associated with Diffserv.

In the preceding example, there is no signaling between the Diffserv network region and network elements outside it. The negotiation of an SLS is the only explicit exchange of resource availability information between the two network regions. ER1 is configured with the information represented by the SLS and as such, is able to act as an admission control agent for the Diffserv network region. Such configuration does not readily support dynamically changing SLSs, since ER1 requires reconfiguration each time the SLS changes. It is also difficult to make efficient use of the resources in the Diffserv network region. This is because admission control does not

consider the availability of resources in the Diffserv network region along the specific path that would be impacted.

By contrast, when the Diffserv network region is RSVP aware, the admission control agent is part of the Diffserv network. As a result, changes in the capacity available in the Diffserv network region can be indicated to the Intserv-capable nodes outside the Diffserv region via RSVP. By including routers interior to the Diffserv network region in RSVP signaling, it is possible to simultaneously improve the efficiency of resource usage within the Diffserv region and to improve the level of confidence that the resources requested at admission control are indeed available at this particular point in time. This is because admission control can be linked to the availability of resources along the specific path that would be impacted. We refer to this benefit of RSVP signaling as "topology aware admission control". A further benefit of supporting RSVP signaling within the Diffserv network region is that it is possible to effect changes in the provisioning of the Diffserv network region (e.g., allocating more or less bandwidth to the EF queue in a router) in response to resource requests from outside of the Diffserv region.

Various mechanisms may be used within the Diffserv network region to support dynamic provisioning and topology aware admission control. These include aggregated RSVP, per-flow RSVP and bandwidth brokers, as described in the following paragraphs.

5.2.1 Aggregated or Tunneled RSVP

A number of drafts [3,6,15, 16] propose mechanisms for extending RSVP to reserve resources for an aggregation of flows between edges of a network. Border routers may interact with core routers and other border routers using aggregated RSVP to reserve resources between edges of the Diffserv network region. Initial reservation levels for each service level may be established between major border routers, based on anticipated traffic patterns. Border routers could trigger changes in reservation levels as a result of the cumulative per-flow RSVP requests from the non-Diffserv regions reaching high or low-water marks.

In this approach, admission of per-flow RSVP requests from nodes outside the Diffserv region would be counted against the appropriate aggregate reservations for the corresponding service level. The size of the aggregate reservations may or may not be dynamically adjusted to deal with the changes in per-flow reservations.

The advantage of this approach is that it offers dynamic, topology aware admission control to the Diffserv network region without requiring the level of RSVP signaling processing that would be required to support per-flow RSVP.

We note that resource management of a Diffserv region using aggregated RSVP is most likely to be feasible only within a single administrative domain, as each domain will probably choose its own mechanism to manage its resources.

5.2.3 Per-flow RSVP

In this approach, described in [3], routers in the Diffserv network region respond to the standard per-flow RSVP signaling originating from the Intserv-capable nodes outside the Diffserv region. This approach provides the benefits of the previous approach (dynamic, topology aware admission control) without requiring aggregated RSVP support. Resources are also used more efficiently as a result of the per-flow admission control. However, the demands on RSVP signaling resources within the Diffserv network region may be significantly higher than in an aggregated RSVP approach.

Note that per-flow RSVP and aggregated RSVP are not mutually exclusive in a single Diffserv region. It is possible to use per-flow RSVP at the edges of the Diffserv region and aggregation only in some "core" region within the Diffserv region.

5.2.4 Granularity of Deployment of RSVP Aware Routers

In 5.2.2 and 5.2.3 some subset of the routers within the Diffserv network is RSVP signaling aware (though traffic control is aggregated as opposed to per-flow). The relative number of routers in the core that participate in RSVP signaling is a provisioning decision that must be made by the network administrator.

In one extreme case, only the border routers participate in RSVP signaling. In this case, either the Diffserv network region must be extremely over-provisioned and therefore, inefficiently used, or else it must be carefully and statically provisioned for limited traffic patterns. The border routers must enforce these patterns.

In the other extreme case, each router in the Diffserv network region might participate in RSVP signaling. In this case, resources can be used with optimal efficiency, but signaling processing requirements and associated overhead increase. As noted above, RSVP

aggregation is one way to limit the signaling overhead at the cost

of some loss of optimality in resource utilization.

It is likely that some network administrators will compromise by enabling RSVP signaling on some subset of routers in the Diffserv network region. These routers will likely represent major traffic switching points with over-provisioned or statically provisioned regions of RSVP unaware routers between them.

5.3 Dynamically Provisioned, Non-RSVP-aware Diffserv Region

Border routers might not use any form of RSVP signaling within the Diffserv network region but might instead use custom protocols to interact with an "oracle". The oracle is a hypothetical agent that has sufficient knowledge of resource availability and network topology to make admission control decisions. The set of RSVP aware routers in the previous two examples can be considered collectively as a form of distributed oracle. In various definitions of the "bandwidth broker" [4], it is able to act as a centralized oracle.

6. Implications of the Framework for Diffserv Network Regions

We have described a framework in which RSVP/Intserv style QoS can be provided across end-to-end paths that include Diffserv network regions. This section discusses some of the implications of this framework for the Diffserv network region.

6.1 Requirements from Diffserv Network Regions

A Diffserv network region must meet the following requirements in order for it to support the framework described in this document.

1. A Diffserv network region must be able to provide support for the standard Intserv QoS services between its border routers. It must be possible to invoke these services by use of standard PHBs within the Diffserv region and appropriate behavior at the edge of the Diffserv region.

2. Diffserv network regions must provide admission control information to their "customer" (non-Diffserv) network regions. This information can be provided by a dynamic protocol or through static service level agreements enforced at the edges of the Diffserv region.

3. Diffserv network regions must be able to pass RSVP messages, in such a manner that they can be recovered at the egress of the Diffserv network region. The Diffserv network region may, but is not required to, process these messages. Mechanisms for transparently carrying RSVP messages across a transit network are described in [3,6,15, 16].

To meet these requirements, additional work is required in the areas of:

1. Mapping Intserv style service specifications to services that can be provided by Diffserv network regions.
2. Definition of the functionality required in network elements to support RSVP signaling with aggregate traffic control (for network elements residing in the Diffserv network region).
3. Definition of mechanisms to efficiently and dynamically provision resources in a Diffserv network region (e.g. aggregated RSVP, tunneling, MPLS, etc.). This might include protocols by which an "oracle" (e.g., a centralized server) conveys information about resource availability within a Diffserv region to border routers. One example of such a mechanism is the so-called "bandwidth broker" proposed in [19, 20, 21].

6.2 Protection of Intserv Traffic from Other Traffic

Network administrators must be able to share resources in the Diffserv network region between three types of traffic:

a. End-to-end Intserv traffic - this is typically traffic associated with quantitative QoS applications. It requires a specific quantity of resources with a high degree of assurance.

b. Non-Intserv traffic. The Diffserv region may allocate resources to traffic that does not make use of Intserv techniques to quantify its requirements, e.g. through the use of static provisioning and SLSs enforced at the edges of the region. Such traffic might be associated with applications whose QoS requirements are not readily quantifiable but which require a "better than best-effort" level of service.

c. All other (best-effort) traffic

These three classes of traffic must be isolated from each other by the appropriate configuration of policers and classifiers at ingress points to the Diffserv network region, and by appropriate provisioning within the Diffserv network region. To provide protection for Intserv traffic in Diffserv regions of the network, we suggest that the DSCPs assigned to such traffic not overlap with the DSCPs assigned to other traffic.

7. Multicast

The use of integrated services over Diffserv networks is significantly more complex for multicast sessions than for unicast sessions. With respect to a multicast connection, each participating

region has a single ingress router and zero, one or several egress routers. The difficulties of multicast are associated with Diffserv regions that contain several egress routers. (Support of multicast functionality outside the Diffserv region is relatively

straightforward since every Intserv-capable router along the multicast tree stores state for each flow.)

Consider the following reference network:

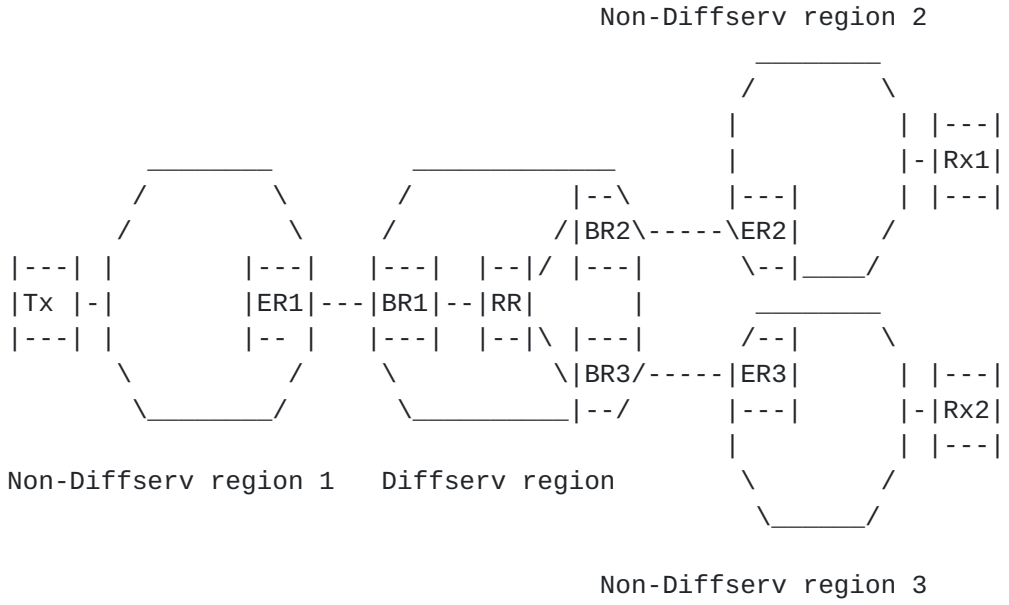


Figure 2: Sample Multicast Network Configuration

The reference network is similar to that of Figure 1. However, in Figure 2, copies of the packets sent by Tx are delivered to several receivers outside of the Diffserv region, namely to Rx1 and Rx2. Moreover, packets are copied within the Diffserv region in a "branch point" router RR. In the reference network BR1 is the ingress router to the Diffserv region whereas BR2 and BR3 are the egress routers.

In the simplest case the receivers, Rx1 and Rx2 in the reference network, require identical reservations. The Diffserv framework [18] supports service level specifications (SLS) from an ingress router to one, some or all of the egress routers. This calls for a "one to many" SLS within the Diffserv region, from BR1 to BR2 and BR3. Given that the SLS is granted by the Diffserv region, the ingress router BR1, or perhaps an upstream node such as ER1, marks packets entering

the Diffserv region with the appropriate DSCP. The packets are routed to the egresses of the Diffserv domain using the original multicast address.

The two major problems, explained in the following, are associated with heterogeneous multicast trees containing branch points within the Diffserv region, i.e. multicast trees where the level of resource requirement is not uniform among all receivers. An example of such a scenario in the network of Figure 2 is the case where both Rx1 and Rx2 need to receive multicast data from Tx1 but only one of

the receivers has requested a level of service above best effort. We consider such scenarios in the following paragraphs.

7.1 Remarking of packets in branch point routers

In the above scenario, the packets that arrive at BR1 are marked with an appropriate DSCP for the requested Intserv service and are sent to RR. Packets arriving at the branch point must be sent towards BR2 with the same DSCP otherwise the service to Rx1 is degraded. However, the packets going from RR towards BR3 need not maintain the high assurance level anymore. They may be demoted to best effort so that the QoS provided to other packets along this branch of the tree is not disrupted. Several problems can be observed in the given scenario:

- In the Diffserv region, DSCP marking is done at edge routers (ingress), whereas a branch point router might be a core router, which does not mark packets.

- Being a core Diffserv router, RR classifies based on aggregate traffic streams (BA), as opposed to per flow (MF) classification. Hence, it does not necessarily have the capability to distinguish those packets which belong to a specific multicast tree and require demotion from the other packets in the behavior aggregate, which carry the same DSCP.

- Since RR may be RSVP-unaware, it may not participate in the admission control process, and would thus not store any per-flow state about the reservations for the multicast tree. Hence, even if RR were able to perform MF classification and DSCP remarking, it would not know enough about downstream reservations to remark the DSCP intelligently.

These problems can be tackled by the following mechanisms:

1. If some Intserv-capable routers are placed within the Diffserv region, it might be possible to administer the network topology and routing parameters so as to ensure that branch points occur only within such routers. These routers would support MF classification and remarking and hold per-flow state for the heterogeneous reservations for which they are the branch point. Note that in this case, branch point routers would have essentially the same functionality as ingress routers of an RSVP-aware Diffserv domain.

2. Packets sent on the "non-reserved" branch (from RR towards BR3) are marked with the "wrong" DSCP; that is, they are not demoted to best effort but retain their DSCP. This in turn requires over reservation of resources along that link or runs the risk of degrading service to packets that legitimately bear the same DSCP along this path. However, it allows the Diffserv routers to remain free of per-flow state.

3. A combination of mechanism 1 and 2 may be an effective compromise. In this case, there are some Intserv-capable routers in the core of the network, but the network cannot be administered so that ALL branch points fall at such routers.

4. Administrators of Diffserv regions may decide not to enable heterogeneous sub-trees in their domains. In the case of different downstream reservations, a ResvErr message would be sent according to the RSVP rules. This is similar to the approach taken for Intserv over IEEE 802 Networks [2,5].

5. In [3], a scheme was introduced whereby branch point routers in the interior of the aggregation region (i.e. the Diffserv region) keep reduced state information regarding the reservations by using measurement based admission control. Under this scheme, packets are tagged by the more knowledgeable Intserv edges routers with scheduling information that is used in place of the detailed Intserv state. If the Diffserv region and branch point routers are designed following that framework, demotion of packets becomes possible.

7.2 Multicast SLSs and Heterogeneous Trees

Multicast flows with heterogeneous reservations present some challenges in the area of SLSs. For example, a common example of an SLS is one where a certain amount of traffic is allowed to enter a Diffserv region marked with a certain DSCP, and such traffic may be destined to any egress router of that region. We call such an SLS a homogeneous, or uniform, SLS. However, in a multicast environment, a single packet that is admitted to the Diffserv region may consume

resources along many paths in the region as it is replicated and forwarded towards many egress routers; alternatively, it may flow along a single path. This situation is further complicated by the possibility described above and depicted in Figure 2, in which a multicast packet might be treated as best effort along some branches while receiving some higher QOS treatment along others. We simply note here that the specification of meaningful SLSs which meet the needs of heterogeneous flows and which can be met be providers is likely to be challenging.

Dynamic SLSs may help to address these issues. For example, by using RSVP to signal the resources that are required along different branches of a multicast tree, it may be possible to more closely approach the goal of allocating appropriate resources only where they are needed rather than overprovisioning or underprovisioning along certain branches of a tree. This is essentially the approach described in [15].

8. Security Considerations

8.1 General RSVP Security

We are proposing that RSVP signaling be used to obtain resources in both Diffserv and non-Diffserv regions of a network. Therefore, all RSVP security considerations apply [9]. In addition, network administrators are expected to protect network resources by configuring secure policers at interfaces with untrusted customers.

8.2 Host Marking

Though it does not mandate host marking of the DSCP, our proposal does allow it. Allowing hosts to set the DSCP directly may alarm network administrators. The obvious concern is that hosts may attempt to "steal" resources. In fact, hosts may attempt to exceed negotiated capacity in Diffserv network regions at a particular service level regardless of whether they invoke this service level directly (by setting the DSCP) or indirectly (by submitting traffic that classifies in an intermediate marking router to a particular DSCP).

In either case, it will generally be necessary for each Diffserv network region to protect its resources by policing to assure that customers do not use more resources than they are entitled to, at each service level (DSCP). The exception to this rule is when the host is known to be trusted, e.g. a server that is under the control

of the network administrators. If an untrusted sending host does not perform DSCP marking, the boundary router (or trusted intermediate routers) must provide MF classification, mark and police. If an untrusted sending host does perform marking, the boundary router needs only to provide BA classification and to police to ensure that the customer is not exceeding the aggregate capacity negotiated for the service level.

In summary, there are no additional security concerns raised by marking the DSCP at the edge of the network since Diffserv providers will have to police at their boundaries anyway. Furthermore, this approach reduces the granularity at which border routers must police, thereby pushing finer grain shaping and policing responsibility to the edges of the network, where it scales better and provides other benefits described in Section 3.3.1. The larger Diffserv network regions are thus focused on the task of protecting their networks, while the Intserv-capable nodes are focused on the task of shaping and policing their own traffic to be in compliance with their negotiated Intserv parameters.

9. Acknowledgments

Authors thank the following individuals for their comments that led to improvements to the previous version(s) of this document: David Oran, Andy Veitch, Curtis Villamizer, Walter Weiss, Francois le Faucheur and Russell White.

Many of the ideas in this document have been previously discussed in the original Intserv architecture document [10].

10. References

[1] Braden, R., Zhang, L., Berson, S., Herzog, S. and Jamin, S., "Resource Reservation Protocol (RSVP) Version 1 Functional Specification", RFC 2205, Proposed Standard, September 1997

[2] Yavatkar, R., Hoffman, D., Bernet, Y., Baker, F. and Speer, M., "SBM (Subnet Bandwidth Manager): A Protocol For RSVP-based Admission Control Over IEEE 802 Style Networks", Internet Draft, draft-ietf-issll-is802-sbm-08.txt, May 1999

[3] Berson, S. and Vincent, R., "Aggregation of Internet Integrated Services State", Internet Draft, draft-berson-rsvp-aggregation-00.txt, August 1998.

[4] Nichols, K., Jacobson, V. and Zhang, L., "A Two-bit

Differentiated Services Architecture for the Internet", RFC 2638, July 1999.

[5] Seaman, M., Smith, A., Crawley, E., and Wroclawski, J., "Integrated Service Mappings on IEEE 802 Networks", Internet Draft, draft-ietf-issll-is802-svc-mapping-04.txt, June 1999

[6] Guerin, R., Blake, S. and Herzog, S., "Aggregating RSVP based QoS Requests", Internet Draft, draft-guerin-aggreg-rsvp-00.txt, November 1997.

[7] Nichols, Kathleen, et al., "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[8] Blake, S., et al., "An Architecture for Differentiated Services." RFC 2475, December 1998.

[9] Baker, F., "RSVP Cryptographic Authentication", Internet Draft, draft-ietf-rsvp-md5-08.txt, February 1999

[10] Braden, R., Clark, D. and Shenker, S., "Integrated Services in the Internet Architecture: an Overview", Internet RFC 1633, June 1994

[11] Garrett, M. W., and Borden, M., "Interoperation of Controlled-Load Service and Guaranteed Service with ATM", RFC2381, August 1998.

[12] Weiss, Walter, Private communication, November 1998.

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

[14] Bernet, Y., "Usage and Format of the DCLASS Object with RSVP Signaling", Internet Draft, draft-issll-dclass-00.txt, August 1999.

[15] Baker, F., Iturralde, C., le Faucheur, F., and Davie, B. "RSVP Reservation Aggregation", Internet Draft, draft-ietf-issll-aggregation-00.txt, September 1999.

[16] Terzis, A., Krawczyk, J., Wroclawski, J., Zhang, L., "RSVP Operation Over IP Tunnels", Internet Draft, draft-ietf-rsvp-tunnel-04.txt, May 1999.

[17] Boyle, J., Cohen, R., Durham, D., Herzog, S., Rajan, D., and Sastry, A., "COPS Usage for RSVP", Internet Draft, draft-ietf-rap-cops-rsvp-05.txt, June 1999.

[18] Bernet, Y., et al., "A Framework for Differentiated Services", Internet draft, draft-ietf-diffserv-framework-02.txt, February 1999.

[19] Jacobson Van, "Differentiated Services Architecture", talk in the Int-Serv WG at the Munich IETF, August 1997.

[20] Jacobson, V., Nichols K., and Zhang, L. "A Two-bit Differentiated Services Architecture for the Internet", RFC 2638, June 1999.

[21] First Internet2 bandwidth broker operability event http://www.merit.edu/internet/working.groups/i2-qbone-bb/inter-op/index.htm

11. Author's Addresses:

Yoram Bernet
Microsoft
One Microsoft Way, Redmond, WA 98052
Phone: (425) 936-9568
Email: yoramb@microsoft.com

Raj Yavatkar
Intel Corporation
JF3-448 2111 NE 25th. Avenue, Hillsboro, OR 97124
Phone: (503) 264-9077
Email: raj.yavatkar@intel.com

Peter Ford
Microsoft
One Microsoft Way, Redmond, WA 98052
Phone: (425) 703-2032
Email: peterf@microsoft.com

Fred Baker

Cisco Systems
519 Lado Drive, Santa Barbara, CA 93111
Phone: (408) 526-4257
Email: fred@cisco.com

Lixia Zhang
UCLA
4531G Boelter Hall Los Angeles, CA 90095
Phone: (310) 825-2695
Email: lixia@cs.ucla.edu

Michael Speer
Sun Microsystems
901 San Antonio Road UMPK15-215 Palo Alto, CA 94303
Phone: (650)786-6368
Email: speer@Eng.Sun.COM

Bob Braden
USC Information Sciences Institute
4676 Admiralty Way Marina del Rey, CA 90292-6695
Phone: (310) 822-1511
Email: braden@isi.edu

Bruce Davie
Cisco Systems
250 Apollo Drive, Chelmsford, MA 01824
Phone: (978) 244-8000
Email: bsd@cisco.com

Eyal Felstaine
Dept. of Computer Science
Technion, IIT
32000 Haifa, Israel
Phone: +972-50-747672
Email: eyalfe@cs.technion.ac.il

John Wroclawski
MIT Laboratory for Computer Science
545 Technology Sq.
Cambridge, MA 02139
Phone: (617) 253-7885
E-mail: jtw@lcs.mit.edu

12. Full Copyright Statement

Copyright (C) The Internet Society (2000). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any

kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."

This draft expires September, 2000