Internet Engineering Task Force
INTERNET-DRAFT

Don Hoffman
Sun Microsystems, Inc.

Raj Yavatkar
Intel Corporation

December, 1996
Expires: June 30, 1997

Integrated-Services/RSVP Requirements for Layer 2 Traffic Control

Status of this Memo

Abstract

This documents discusses some of the requirements placed on L2 traffic control in IEEE 802-style networks by IP Integrated Services (IntServ) and RSVP signaling. It outlines some of the features of IntServ/RSVP which are of particular relevance to this style of network, and defines the of the requirements that a L2 network must meet to fully support these features. Finally, it discusses certain L2 mechanism which may aid in meeting these requirements.

0) What's Changed Since Last Version

1) First version of document.

1) Introduction

This documents discusses some of the requirements placed on L2 traffic control in IEEE 802-style networks by IP Integrated Services (IntServ) and RSVP signaling. It is intended to be used as a condensed set of guidelines for L2 systems architects wishing to provide complete support for an IntServ/RSVP infrastructure.

For the purposes of discussion, this document hypothesizes various L2 mechanisms (especially Section 5). These L2 mechanisms are NOT part of the specification of requirements, and should be taken as suggestions only. This document reflects only the opinions of the authors, and is intended to be used as a starting point for further discussion within the IETF and IEEE 802 communities.

2) Reference environment

For the purposes of the discussion that follows, we make certain assumptions about the reference environment. No claim is made that these are the only valid set of assumptions, and they may be modified after further discussion.

First, we assume an IP-based environment that makes use of the following protocols/specifications in their referenced versions:

IP multicast and unicast datagram service [Ref needed]
RSVP signaling [6]
Integrated Controlled Load Service and Guaranteed Service [1, 2, 3, 4, 5]

We further assume that each IP subnet corresponds to a single L2 "domain". A domain is arbitrarily defined to be the set of nodes and links interconnected without passing through some sort of IP (L3) forwarding function. If new 802-proposed mechanisms such as Virtual LANS (VLANs) are employed, then multiple IP subnets (and multiple L2 domains) could reside on a single physical L2 topology. An "edge-device" (ED) is defined to be an IP network element that is a source of or sink for IP traffic for the subnet. An ED can be either a host or a router. Each ED's interface to the L2 network has a unique IP address. Multiple physical interfaces with the same address (as might be used for L2 load balancing) are considered to be a single interface for the purposes of this discussion. Multiple "virtual interfaces" on the same physical interface are considered to be distinct if they have distinct IP addresses.

For the purposes of this specification, we will model the L2 domain as a

black box, defined as a single logical link with multiple input and output ports. The behavior of this link is assumed to be more complex than strictly FIFO.

The following assumptions are defined in order to clarify the intended scope of this document:

A L2 domain will be built from an arbitrarily large set of L2 elements such as hubs, switches and shared LAN segments. The number of elements that make up a subnet is allowed to be quite large, and the number of EDs also quite large (say an entire class B address).

On certain elements, the maximum possible aggregate input bandwidth may exceed the capacity of any single output port. Similarly, certain L2 links may be shared, and the total possible output bandwidth of all attached devices on to the link may exceed the capacity of the link and any of the input ports of the attached devices.

Some of the L2 elements (e.g, switches and bridges) may buffer data on an output port when the offered load exceeds the rate of that port. The amount of buffering on switches is considered to be finite and usually small.

A bridge or switch may have varying amounts of intelligence in terms of policing and outgoing queue management. This is, for the most part, considered to be less than the average ED and may be none for certain legacy or very low cost devices.

Although all links that make up the L2 topology are considered to be very high speed, their speeds can range over several orders of magnitude. EDs are also considered to support interfaces and transfer rates ranging over several orders of magnitude.

3) Integrated Services and RSVP Background

Certain key feature of RSVP and the IP Integrated Services architecture are described here as they are believed to be important for a full understanding of L3 requirements.

The relevant IETF RFCs and Internet Drafts [1, 2, 3, 4, 5, 6] provide a full description of IP Integrated Services and RSVP and this section is derived from these specifications.

Readers already familiar with the above documents may wish to skip to the next section.

3.1) RSVP

3.1.1) Definition of a flow: The RSVP specification defines a flow as a tuple consisting of the IP destination and source address, an IP protocol ID, and if the protocol is TCP or UDP, the destination and source port number. All but the destination IP address may be defined as a wild-card according to specific rules. Note that a particular IP source and destination address

pair may have several flows (each with different flow specifications distinguished by port number) running between them. Also, there may be best effort traffic between these two nodes not associated with any flow.

A flow is described by a Tspec and an Rspec. The Tspec describes the nature of the flow coming from the sender, and takes the form of a token bucket specification (r, b) plus a peak rate (p), a minimum policed unit (m) and a maximum packet size (M). It is common to all service types. The Rspec described the required QoS from the receiver, and is specific to the service type (e.g., Controlled Load or Guaranteed Service). A reservation request from a receiver will contain both a TSpec component and an RSpec component.

3.1.2) Heterogeneous reservations: Each receiver determines its own flow requirements, and the reservation request from each receiver is free to define a different set of requirements. There are a few restrictions preventing receivers from using different >styles< of reservations, but in general each receiver is free to set any of the parameters of the TSpec or RSpec to receiver-specific values.

Note - reservations can differ not only in the TSpec information, but also in the RSpec information (e.g., max delay in the IntServ Guaranteed Service).

One form of heterogeneity that will almost always be seen is between receivers that have obtained reservations and those which are satisfied with best effort service (or who have not yet requested the reservation).

3.1.3) Policing/reshaping at merge points: In the case of multicast flows, reservations from multiple receivers are, depending on the style of reservation, "merged" at IP multicast branch points as the reservation propagates back up toward the sender. It is then possible that some of the outgoing interfaces at this downstream branch point will not be able to support the full combined flowspec from upstream.

For example, a reservation arrives with a TSpec "r" value of 64Kbps on a 128Kbps link. Another reservation, with an "r" value of 1Mbps, for the same flow and sender arrives on an ethernet interface on the same router. Admission control on each interface can succeed, and a merged reservation of 1Mbps is forwarded toward the sender. A sending ED may offer up to 1Mbps of load toward both the 1Mbps interface and the 128Kbps interface. The slower interface must police this flow to 64Kbps in order to minimize the effects on other traffic on the interface (both reserved and non-reserved). (See Section **3.2 discussion on handling of excess traffic.)**

In addition, a reservation may be "shared" among multiple senders (in the case of WF or SE reservation styles). In such cases, the total possible aggregate offered load from all the senders may exceed the reservation on a single outgoing interface by a significant amount. In certain conferencing applications this can be by a factor of several hundred. The application

assumes in this case that some external mechanisms (which may not always be reliable) prevents too many senders from transmitting at once.

Finally, the effects of queuing at intermediate systems may cause sufficient traffic rate distortion that a compliant flow no longer remains compliant with the Tspec.

Because of these three scenarios, the RSVP/IntServ architecture requires that reshaping and/or policing be done at all source merge points and at heterogeneous branch points. These policing points are known to RSVP and provided to local traffic control mechanisms on each outgoing interface.

(Side note - It should be noted, however, that the more extreme cases discussed above will not be common in actual operation.)

3.1.4) Scalability: One motivation for RSVP's receiver-orientation is to achieve very large scale multicast fan-out. A key part of this is the merging process mentioned above. It is still uncertain how RSVP will scale on subnets with VERY large fanout within a single hop (many of thousands, as might be seen in a single campus wide L2 topology), where the merging functions are of limited assistance. Although the soft-state refresh interval for RESV messages can be set arbitrarily long, this is in conflict with responsive recovery from certain error conditions.

3.2) Integrated Services

3.2.1) Controlled Load Service: The basic behavior of a compliant Controlled Load (CL) Service stream is approximated by the behavior visible to applications receiving best-effort service under unloaded conditions. This behavior should be seen by all compliant CL streams even in the case of severe congestion for best-effort traffic. The implication of this is that congestion in the best-effort class should not interfere with CL traffic. This can generally be taken to imply that some sort of priority, traffic limiting or traffic separation scheme should be implemented on each outgoing interface, and if the link is multi-access (i.e., multiple senders), an equivalent scheme should be implemented on the link (or coordinated across all output ports on to the link).

The CL specification requires that if a flow is non-conformant to the TSpec, the forwarding node MUST attempt to forward excess traffic on a best-effort basis. Further, non-conformance will not be unusual at merge and branch points and will happen as "a matter of normal operation." Non-conformant traffic must not interfere with conformant CL traffic in other flows.

The specification suggests that a marking scheme be used for non-conformant traffic if one is available.

In the case of flow non-conformance, the forwarding element is allowed to degrade all the flow's packets equally, or it may sort the flow's traffic into conformant and non-conformant sets. In the latter case the packets in a flow may be reordered by the network elements.

3.2.2) Guaranteed Service: The basic behavior of a Guaranteed Service (GS) flow is an assured level of bandwidth that produces a delay-bounded service with no queuing loss for all conforming datagrams. Unlike CL service, there

are fairly specific requirements on the behavior of all forwarding elements in the path. Briefly, the end-end behavior conforms to the fluid model within a specified set of error bounds. The GS draft should be consulted for a full understanding of these requirements.

As with CL services, traffic is is policed at the edge of the network and non-conforming traffic should be treated as best-effort datagrams (with the same implications with regard to packet reordering). The handling at interior merge and branch point is different, however. In this case, non-conforming flows are "reshaped" by delaying datagrams until the flow is within conformance of the TSpec. Again, reordering is allowed, but the GS specification suggests ways that this can be minimized.

4) Layer 2 Requirements

Unless otherwise specified, the requirements defined in this section are mandatory for a L2 mapping of IP Integrated Services to be considered compliant. (Note - These are offered for discussion, and my be seriously modified in future versions of the draft)

Unless otherwise specified, a compliant L3/L2 mapping must maintain the RSVP and Integrated Services semantics and behavior defined in section 3 for the services it supports (at this time either Controlled Load or Guaranteed Service). This includes semantics not mentioned directly, but covered in referenced specifications. No specific L2 or L3 mechanisms to accomplish this are required by this document.

Taken as a black box, the L2 domain can be considered to be a single shared link with multiple input and output interfaces. As such, it can also be considered an implicit merge and split point. The behavior of this link is assumed to be more complex than strictly FIFO. As they are based on a black box model, these requirements do not mandate policing and reshaping in the interior of the L2 domain. (Although it may be desirable. See Section 5.) Nor do they mandate any particular scheduling algorithm.

Instead, compliance is measured at the receiving ED, based on IS-compliant behavior at the sending EDs. An L2 traffic control mechanism is considered compliant if the traffic, measured at a point after each receiving ED's input queue, meets the Rspec for the flow, assuming:

> The flow sources are policed/reshaped at each ED output interface onto the domain according to the appropriate TSpec for that flow.
>
> The aggregate flow bandwidth from all sender ED output interface does not exceed the merged Tspec at any particular receiver. Merging is defined to be according to the reservation style in use for that flow.

If one of the above assumptions is not met, compliant behavior for the L2 traffic control mechanism is not defined. In particular, the second assumption may be violated in the course of normal operation. Several complexities arise from this:

> The flow may be multicast, and the available bandwidth on each of the receiver input ports may be different (For example, some receivers have 100Mbps input ports, other only have 10Mbps input ports). Consequently, it might be possible for all assumptions to apply to only a subset of the receivers. In this case, a traffic control mechanism is considered compliant if the flow is compliant with the RSpec at the subset of receivers that satisfy all assumptions.
>
> These requirements do not mandate that the effective utilized bandwidth at ED input ports be less than the merged (were appropriate) TSpecs for shared reservation styles. But, a compliant L2 traffic control environment MUST continue to meet the Rspec for other flows where the above input assumptions are met. For example, in cases where many senders to a WF flow all send at once, exceeding the merged Tspec at all receivers, compliant behavior for that flow is undefined, and may result in violating the Rspec. A compliant L2 traffic control implementation will not cause other flows that do meet the input assumptions to violate the Rspec.

An L2 traffic control mechanism may (and probably will) provide some way to limit the total number of reserved flows, or the total amount of bandwidth allowed in that domain. The exact nature of that mechanism and the interface between L3 signaling and the L2 mechanism is outside the scope of this document.

As multicast is a key element of many of the applications that make use of Integrated Services, the mechanisms provided in L2 traffic control must scale to the maximum number of nodes anticipated for that domain.

5) L2 Pragmatics

The approaches discussed in this section are not firm requirements, but are to be taken as suggestions for possible mechanisms for implementing the Integrated Services and RSVP functions over 802-style networks. The implementor or specified of L2 mechanisms is free to employ other approaches. Multiple mechanisms may be suggested in several cases. (Note - These are offered for discussion, and my be seriously modified or dropped in future versions of the draft.)

5.1) Policing/reshaping at merge/split points.

As mentioned above, a L2 domain is an implicit merge/split point for RSVP

flows. As long as the requirements in section 4 are met, there is no specific requirement to do policing or reshaping at these implicit merge

points. It is likely, however, that in order to avoid congestion of internal links and ED input ports some sort of policing/reshaping may be desirable internal to the L2 switched environment. Several mechanisms may be possible:

The L2 environment may employ sufficiently conservative admission control criteria such that offered loads significantly over the receiver TSpec do not result in congestion or delay bound violation. No policing/reshaping is done.

Police/reshape on aggregations of flows. Note that policing mechanisms based on aggregated sets of flows may result in service degradation for conformant flows due to non-conformance by other flows in the same aggregation.

Implement IntServ-style per-flow mechanisms in L2.

The first two approaches may result in behavior that does not meet the requirements in Section 4. This will be the most problematic for GS, but may provide a useful approximation of CL service in certain environments. (See Section 5.3.)

5.2) Flow identification

It is probably undesirable to require flow packet classification based on IP header information in all L2 elements. The L2 mechanism may use some sort of information in the L2 header (e.g., VLAN tag, flow tag, COS field or priority bits) to facilitate packet processing in the interior of the L2 domain. In this case, the supplemental L2 header information may be derived based on information provided by the L3 ED, obtained from the IP flow classifier in that node.

If full per-flow policing and merging is implemented in the interior of the L2 domain, then the L2 header info must key to a specific IP flow. If no or only loose policing is done then this header information may map to some aggregation of IP flows (e.g., based on service type).

Note that policing mechanisms based on aggregated sets of flows may result in service degradation for conformant flows due to non-conformance by other flows.

5.3) Service approximation

In the above sections, several cases are discussed where extreme cases of receiver heterogeneity and sender fan-in can result in significant issues for a compliant L2 traffic control mechanism. It may be worthwhile to define approximations to full compliance that meet the practical requirements of actual applications in real-life situations. Future versions of this document may talk more specifically on agreed-on approximations.

Expires: June 30, 1997

References:

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

[2] S. Shenker and J. Wroclawski. "Network Element QoS Control Service Specification Template". Internet Draft, July 1996, <draft-ietf-intserv-svc-template-03.txt>

[3] J. Wroclawski, "Specification of the Controlled-Load Network Element Service", Internet Draft, August 1996, <draft-ietf-intserv-ctrl-load-svc-03.txt>

[4] S. Shenker et. al., "Specification of Guaranteed Quality of Service", Internet Draft, August 1996, <draft-ietf-intserv-guaranteed-svc-06.txt>

[5] J. Wroclawski, "Use of RSVP with IETF Integrated Services", Internet Draft, July 1996, <draft-ietf-intserv-rsvp-use-00.txt>

[6] B. Braden, et. al. "Resource Reservation Protocol (RSVP) - Version 1 Functional Specification", Internet Draft, July 1996, <draft-ietf-rsvp-spec-13.txt>

Authors' Addresses:

Don Hoffman
Sun Microsystems, Inc.
MS: UMPK14-305
2550 Garcia Avenue
Mountain View, California 94043-1100
USA
phone: +1 503-297-1580
email: don.hoffman@eng.sun.com

Raj Yavatkar
Intel Corporation
MS: JF3-206
2111 N.E. 25th Avenue,
Hillsboro, OR 97124
USA

phone: +1 503-264-9077
email: yavatkar@ibeam.intel.com