Network Working Group
Request for Comments: 2688
Category: Standards Track

S. Jackowski
Deterministic Networks
D. Putzolu
Intel Architecture Labs
E. Crawley
Argon Networks
B. Davie
Cisco Systems

September 1999

Integrated Services Mappings for Low Speed Networks

Status of this Memo

Copyright Notice

Abstract

A set of companion documents describe an architecture for providing integrated services over low-bitrate links, such as modem lines, ISDN B-channels, and sub-T1 links [1, 2, 3, 4]. The main components of the architecture are: a set of real-time encapsulation formats for asynchronous and synchronous low-bitrate links, a header compression architecture optimized for real-time flows, elements of negotiation protocols used between routers (or between hosts and routers), and announcement protocols used by applications to allow this negotiation to take place.

This document defines the service mappings of the IETF Integrated Services for low-bitrate links, specifically the controlled load [5] and guaranteed [6] services. The approach takes the form of a set of guidelines and considerations for implementing these services, along with evaluation criteria for elements providing these services.

1. Introduction

In addition to the "best-effort" services the Internet is well-known for, other types of services ("integrated services") are being developed and deployed in the Internet. These services support special handling of traffic based on bandwidth, latency, and other requirements that cannot usually be met using "best-effort" service.

This document defines the mapping of integrated services "controlled load" [5] and "guaranteed" [6] services on to low-bandwidth links. The architecture and mechanisms used to implement these services on such links are defined in a set of companion documents. The mechanisms defined in these documents include both compression of flows (for bandwidth savings) [4,10] and a set of extensions to the PPP protocol which permit fragmentation [2] or suspension [3] of large packets in favor of packets from flows with more stringent service requirements.

1.1. Specification Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [11].

2. Issues for Providing Controlled and Guaranteed Service

Unlike other link layers, the links referred to in this document operate only over low speed point to point connections. Examples of the kinds of links addressed here include dial-up lines, ISDN channels, and low-speed (1.5Mbps or less) leased lines. Such links can occur at different positions within the end-to-end path:

- host to directly connected host.
- host to/from network access device (router or switch).
- Edge device (subnet router or switch) to/from router or switch.
- In rare circumstances, a link from backbone router to backbone router.

These links often represent the first or last wide area hop in a true end to end service. Note that these links may be the most bandwidth constrained along the path between two hosts.

The services utilized in mapping integrated services to these links are only provided if both endpoints on the link support the architecture and mechanisms referenced above. Support for these mechanisms is determined during the PPP negotiation. The non-shared

nature of these links, along with the fact that point-to-point links are typically dual simplex (i.e., the send and receive channels are separate) allows all admission control decisions to be made locally.

As described in [2] and [3], for systems that can exert real time control of their transmission at a finer grain than entire HDLC frames, the suspend/resume approach optimizes the available bandwidth by minimizing header overhead associated with MLPPP pre-fragmentation and can provide better delay. However, this comes at the expense of preparing all outgoing data and scanning all incoming data for suspend/resume control information. The fragmentation approach can be implemented without additional scanning of the data stream (beyond bit-/byte-stuffing, which may be in hardware) and is applicable to systems which provide only frame-oriented transmission control. Choice of suspend/resume versus fragmentation should be made based on the level of transmission control, the element's capability to handle the HDLC-like framing described in [2], and the system overhead associated with byte by byte scanning (required by suspend/resume).

To provide controlled load or guaranteed service with the suspend/resume approach, when a packet for an admitted flow (QoS packet) arrives during transmission of a best effort packet and continued transmission of the best effort packet would violate delay constraints of the QoS service flows, the best effort packet is preempted, the QoS packet/fragments are added to the transmission, and the best effort packet transmission is then resumed: usually all in one transmission. The receiving station separates the best effort packet from the embedded QoS packet's fragments. It is also conceivable that one QoS flow's packet might suspend another flow's packet if the delivery deadline of the new packet is earlier than the current packet.

For systems which use fragmentation, any packets longer than the maximum tolerable delay for packets from enhanced service flows are fragmented prior to transmission so that a short packet for another flow can be interleaved between fragments of a larger packet and still meet the transmission deadline for the flow requiring enhanced services.

Note that the fragmentation discussed in this document refers to multilink PPP (MLPPP) fragmentation and associated MCMLPPP modifications as described in [2], not IP or other layer 3 fragmentation. MLPPP fragmentation is local to the PPP link, and does not affect end-to-end (IP) MTU.

2.1 Calculating "Acceptable Delay" for Int-serv flows

A router which provides Controlled Load or Guaranteed Service over a low speed serial link needs to have some notion of the "acceptable delay" for packets that belong to int-serv flows. If using fragmentation, a router needs to know what size to fragment packets to; if using suspend/resume, it needs to know when it is appropriate to suspend one packet to meet the delay goals of another.

Unfortunately, there is no hard and fast way for a single delay bound to be determined for a particular flow; while the end-points of a flow have enough information to determine acceptable end-to-end delay bounds and to make reservation requests of the network to meet those bounds, they do not communicate a "per-hop" delay to routers.

In the case of Guaranteed Service [6], one approach is to let the network operator configure parameters on the router that will directly affect its delay performance. We observe that guaranteed service allows routers to deviate from the ideal fluid flow model and to advertise the extent of the deviation using two error terms C and D, the rate-dependent and rate-independent error terms, defined in [6]. A network operator can configure parameters of the low speed link in such a way that D is set to a value of her choice.

If link-level fragmentation is used, the router controlling a low-speed link can be configured with a certain fragment size. This will enable a component of the error term D to be calculated based on the time to send one fragment over the link. (Note that D may have other components such as the speed of light delay over the link.) Details of the calculation of D are described below. Similarly, if suspend/resume is used, the router may be configured with a delay parameter, which would enable it to decide when it was appropriate to suspend a packet.

For Controlled Load, there are no error terms, and the router must decide how best to meet the requirements of the admitted reservations using only the information in their TSpecs. Since the definition of Controlled Load states that a CL flow with Tspec rate r should receive treatment similar to an unloaded network of capacity r, CL packets should not generally experience end-to-end delays significantly greater than b/r + propagation delays. Clearly a router connected to a low speed link should not introduce a delay greater than b/r due to transmission of other fragments; ideally it should introduce substantially less delay than b/r, since other hops on the end-to-end path may introduce delay as well. However, this may be difficult for flows with very small values of b.

It is expected that implementers will make their own tradeoffs as to how low to make the delay for Controlled Load flows. Similarly, it may not be possible or desirable to configure the parameters affecting D to arbitrarily small values, since there is a cost in overhead in fragmenting packets to very small sizes. Conversely, if D is too large, some applications may find that they cannot make a reservation that will meet their delay objectives.

For the remainder of this document, we assume that a router has some notion of the acceptable delay that it may introduce before beginning transmission of a packet. This delay is in addition to any delay that a packet might be subjected to as a result of the "ideal" queuing algorithm that the router uses to schedule packets.

3. Controlled Load and Guaranteed Service Class Mapping

Supporting integrated services over PPP links which implement MCML or RTF can be accomplished in several ways. Guidelines for mapping these services to PPP links and to the classes provided by the suspend/resume and fragmentation mechanisms are presented below. Note that these guidelines assume that some sort of signaling protocol is used to indicate desired quality of service to both the sender and receiver of a flow over a PPP link.

3.1 Predefined Class Mappings

A relatively simple method of class mapping that MAY be used is one where class values correspond to predefined levels of service. In this arrangement, all admitted flows are grouped into one of several buckets, where each bucket roughly corresponds to the level of service desired for the flows placed in it. An example set of mappings appears below:

MCML Short	MCML Long	RTF	Service
0b00	0b0000	0b000	Best Effort
NA	0b0001	0b001	Reserved
0b01	0b0010	0b010	Delay Sensitive, no bound
NA	0b0011	0b011	Reserved
NA	0b0100	0b100	Reserved
0b10	0b0101	0b101	Delay Sensitive, 500ms bound
NA	0b0110	0b110	Delay Sensitive, 250ms bound
0b11	0b0111	0b111	Network Control

Table 1: Example Mappings of Classes to Services

Note that MCML has two formats, short sequence numbers, and long sequence numbers, that allow for 2 and 4 bits of class identification. RTF allows for 3 bits of class identification in all formats.

Using a default-mapping method of assigning classes to flows in a fixed fashion comes with certain limitations. In particular, all flows which fall within a particular bucket (are assigned to a particular class) will be scheduled against each other at the granularity of packets, rather than at the finer grained level of fragments. This can result in overly conservative admission control when the number of available classes is small such as in MCML short sequence number format.

3.2 Predefined Class Mappings and Prefix Elision

In the case where fewer reservations are expected than the total number of classes negotiated for a PPP link, it is possible to assign individual flows to fixed class numbers. This assignment is useful in the case where the protocol identifier associated with one or more flows is known at LCP negotiation time and the bandwidth of the connection is relatively small. If these conditions hold true, then for those flows that are known, a specific class can optionally be assigned to them and the prefix elision PPP option [2] can be used for those classes to achieve a small bandwidth savings.

3.3 Dynamic Class Mappings

In the case where predefined class mappings are not satisfactory, an implementer MAY map class values to individual packets rather than assigning flows to fixed classes. This can be done due to the fact that the classes that MCML and RTF provide can be viewed purely as PPP-specific segmentation/fragmentation mechanisms. That is, while the class number MUST remain constant on an intra-packet basis, it MAY vary on an inter-packet basis for all flows transiting a PPP link. Actual assignment of particular flows to fixed classes is unnecessary, as the class numbers are NOT REQUIRED to have any meaning other than in the context of identifying the membership of fragments/segments as part of a single packet. This point is sufficiently important that an example is provided below.

Consider a PPP link using the MCML short sequence number fragment format (that is, four classes are provided). Assume that in addition to carrying best effort traffic, this link is carrying five guaranteed service flows, A, B, C, D, and E. Further assume that the link capacity is 100kbit/s and the latency is 100ms. Finally, assume the BE traffic is sufficient to keep the pipe full at all times and that GS flows A-E are each 10kbit/s and all have delay bounds of 145ms.

```
Time(ms)        Action
 0     BE traffic is queued up
 0     2kbit fragment from 10kbit packet of BE traffic sent, cls 0 (...)
 8     2kbit fragment from BE sent, cls 0 (10kbit BE packet done)
 9     8kbit packet from flow A arrives
10     2kbit fragment from A sent, cls 1 (8kbit flow A packet start)
11     8kbit packet from flow B arrives
12     2kbit fragment from B sent, cls 2 (8kbit flow B packet start)
13     8kbit packets from flows C, D, and E arrive
14     2kbit fragment from C sent, cls 3 (8kbit flow C packet start)
16     2kbit fragment from D sent, cls 0 (8kbit flow D packet start)
18     2kbit fragment from A sent, cls 1
20     2kbit fragment from B sent, cls 2
22     2kbit fragment from A sent, cls 1
24     2kbit fragment from A sent, cls 1 (8kbit flow A packet done)
26     2kbit fragment from E sent, cls 1 (8kbit flow E packet start)
27     8kbit packet from flow A arrives
28     2kbit fragment from B sent, cls 2
30     2kbit fragment from C sent, cls 3
32     2kbit fragment from E sent, cls 1
34     2kbit fragment from B sent, cls 2 (8kbit flow B packet done)
36     2kbit fragment from E sent, cls 1
38     2kbit fragment flow A sent, cls 2 (8kbit flow A packet start)
       (etc.)
```

This example shows several things. First, multiple flows MAY share the same class, particularly in the case where there are more flows than classes. More importantly, there is no reason that a particular flow must be assigned to a fixed class - the only requirement is that each packet, when fragmented, MUST have the same class value assigned to all fragments. Beyond this requirement the link scheduler may assign individual to changing class numbers as necessary to meet reservation requirements.

One suggestion to implementers of integrated services on MCML and RTF links using dynamic mappings is that all BE traffic SHOULD be logically separated from QoS traffic, and mapped to a fragmentable (MCML classes 0-3 in short sequence number fragment format, 0-15 in long sequence number fragment format) or suspendable (RTF classes 0-6) class. Since BE traffic will in most implementations not be scheduled for transmission except when a link is empty (that is, no CL or GS traffic is ready for transmission), implementers MAY choose to make use of class number 0 for BE traffic.

3.4 Non-Conformant Traffic

Treatment of non-conformant QoS traffic is largely determined by the appropriate service specifications, but the detailed implementation in the context of this draft allows for some flexibility. Policing of flows containing non-conformant traffic SHOULD always be done at the level of granularity of individual packets rather than at a finer grained level. In particular, in those cases where a network element scheduling flows for transmission needs to drop non-conformant traffic, it SHOULD drop entire packets rather than dropping individual fragments of packets belonging to non-conformant traffic. In those cases where a network element forwards non-conformant traffic when link bandwidth is available rather than dropping the traffic, the implementation SHOULD fragment packets of such traffic as if it were best effort traffic.

Whether BE and non-conformant traffic are treated differently in regards to transmission (e.g., BE is given priority access over non-conformant traffic to the link) or whether within each type of traffic special treatment is afforded to individual flows (e.g., WFQ, RED, etc.) is service dependent.

4. Guidelines for Implementers

4.1. PPP Bit and Byte Stuffing Effects on Admission Control

An important consideration in performing admission control for PPP links is reductions in effective link rate due to bit stuffing. Typical bit stuffing algorithms can result in as much as 20% additional overhead. Thus, admission control implementations for guaranteed service over links where bit stuffing is used SHOULD take the RSpec rate of all flows and multiply by 1.2, to account for the 20% overhead from bit stuffing, when determining whether a new flow can be admitted or not. Admission control implementations for controlled load reservations may use a similar algorithm using the TSpec peak rate or may attempt to measure the actual degree of expansion occurring on a link due to bit stuffing. This characterization can then be used to adjust the calculated remaining link capacity. Such measurements must be used cautiously, in that the degree of bit stuffing that occurs may vary significantly, both in an inter- and intra-flow fashion.

Byte stuffing is also used on many PPP links, most frequently on POTS modems when using the v.42 protocol. Byte stuffing poses a difficult problem to admission control, particularly in the case of guaranteed service, due to its highly variable nature. In the worse case, byte stuffing can result in a doubling of frame sizes. As a consequence, a strict implementation of admission control for guaranteed load on

byte stuffed PPP links SHOULD double the RSpec of link traffic in making flow admission decisions. As with bit stuffing, implementations of controlled load service admission control algorithms for links with byte stuffing MAY attempt to determine average packet expansion via observation or MAY use the theoretical worst case values.

4.2. Compression Considerations

The architecture for providing integrated services over low bandwidth links uses several PPP options to negotiate link configuration as described in [4, 8, 10]. When deciding whether to admit a flow, admission control MUST compute the impact of the following on MTU size, rate, and fragment size:

Header compression: Van Jacobson or Casner-Jacobson [4,8,10].
Prefix Elision.
CCP.
Fragment header option used.
Fragmentation versus suspend/resume approach.

If any of the compression options are implemented for the connection, the actual transmission rate, and thus the bandwidth required of the link, will be reduced by the compression method(s) used.

Prefix elision can take advantage of mapping flows to MLPPP classes to elide prefixes which cannot be compressed at higher layers. By establishing agreement across the link, the sender may elide a prefix for a certain class of traffic and upon receiving packets in that class, the receiver can restore the prefix.

Both compression gain and elision gain MUST be included as described in the admission control section below. Note that the ability to perform compression at higher layers (e.g. TCP or RTP/UDP) may depend on the provision of a hint by the sender, as described in [9].

4.3. Admission Control

Admission control MUST decide whether to admit a flow based on rate and delay. Assume the following:

LinkRate is the rate of the link.
MTU is the maximum transmission unit from a protocol.
MRU is the maximum receive unit for a particular link.
CMTU is the maximum size of the MTU after compression is applied.
eMTU is the effective size at the link layer of an MTU-sized packet after link layer fragmentation and addition of the fragment headers.
FRAG is the fragment size including MLPPP header/trailers.

Header is the size of the header/trailers/framing for MLPPP/Fragments.
pHeader is the additional header/framing overhead associated with suspend/resume. This should include FSE and worst case stuffing overhead.
pDelay is the time take to suspend a packet already "in flight", e.g. due to the delay to empty the output FIFO.
b is the bucket depth in bytes
R is the requested Rate.
Dlink is the fixed overhead delay for the link (Modem, DSU, speed-of-light, etc).
eRate is the effective rate after compression and fragmentation.

The Dlink term MAY be configured by an administrative tool once the network is installed; it may be determined by real-time measurement means; or it MAY be available from hardware during link setup and/or PPP negotiation. Refer to Appendix A for more considerations on PPP link characteristics and delays.

Admission control MUST compute CMTU, eMTU, and eRate for Controlled Load Service, and it MUST compute CMTU, eMTU, eRate, and D for Guaranteed Service:

To determine whether the requested rate is available, Admission Control MUST compute the effective rate of the request (eRate) - worst case - as follows:

```
#_of_Fragments = CMTU div (FRAG-Header) [Integer divide]
Last_Frag_Size = CMTU mod (FRAG-Header

If Last_Frag_Size != 0
   eMTU = (#_of_Fragments) * FRAG + Last_Frag_Size + Header
Else
   eMTU = (#_of_Fragments) * FRAG

eRate = eMTU/CMTU * R [floating point divide]
```

Admission control SHOULD compare the eRate of the request against the remaining bandwidth available to determine if the requested rate can be delivered.

For Controlled Load Service, a flow can be admitted as long as there is sufficient bandwidth available (after the above computation) to meet the rate requirement, and if there is sufficient buffer space (sum of the token bucket sizes does not exceed the buffer capacity). While some statistical multiplexing could be done in computing admissibility, the nature of the low-bitrate links could make this approach risky as any delay incurred to address a temporary overcommitment could be difficult to amortize.

4.4 Error Term Calculations

Guaranteed Service requires the calculation of C and D error terms. C is a rate-dependent error term and there are no special considerations affecting its calculation in the low-speed link environment. The D term is calculated from the inherent link delay (Dlink) plus the potential worst case delay due to transmission of another fragment or suspend/resume overhead. Thus, D should be calculated as

D = Dlink + FRAG/LinkRate

in the case of a fragementing implementation and

D = Dlink + pHeader + pDelay

for a suspend/resume implementation.

4.5 Scheduling Considerations

We may think of the link scheduler as having two parts, the first of which schedules packets for transmission before passing them to the second part of the scheduler -- the link level scheduler -- which is responsible for fragmenting packets, mapping them to classes, and scheduling among the classes.

In the dynamic class mapping mode of Section 3.3, when deciding which class to assign a packet to, the link level scheduler should take account of the sizes of other packets currently assigned to the same class. In particular, packets with the tightest delay constraints should not be assigned to classes for which relatively large packets are in the process of being transmitted.

In either the dynamic or the static class mapping approach, note that the link-level scheduler SHOULD control how much link bandwidth is assigned to each class at any instant. The scheduler should assign bandwidth to a class according to the bandwidth reserved for the sum of all flows which currently have packets assigned to the class. Note that in the example of Section 3.3, when packets from flows A and E were assigned to the same class (class 1), the scheduler assigned more bandwidth to class 1, reflecting the fact that it was carrying traffic from reservations totaling 20kbit/s while the other classes were carrying only 10kbit/s.

5. Security Considerations

General security considerations for MLPPP and PPP links are addressed in RFC 1990 [12] and RFC 1661 [13], respectively. Security considerations relevant to RSVP, used as the signaling protocol for integrated services, are discussed in RFC 2209 [14].

A specific security consideration relevant to providing quality of service over PPP links appears when relying on either observed or theoretical average packet expansion during admission control due to bit- or byte-stuffing. Implementations based on these packet-expansion values contain a potential vulnerability to denial of service attacks. An adversary could intentionally send traffic that will result in worst case bit- or byte stuffing packet expansion. This in turn could result in quality of service guarantees not being met for other flows due to overly permissive admission control. This potential denial of service attack argues strongly for using a worst case expansion factor in admission control calculations, even for controlled load service.

Beyond the considerations documented above, this document introduces no new security issues on top of those discussed in the companion ISSLL documents [1], [2] and [3] and AVT document [4]. Any use of these service mappings assumes that all requests for service are authenticated appropriately.

6. References

[1] Bormann, C., "Providing Integrated Services over Low-bitrate Links", RFC 2689, September 1999.

[2] Bormann, C., "The Multi-Class Extension to Multi-Link PPP", RFC 2686, September 1999.

[3] Bormann, C., "PPP in a Real-time Oriented HDLC-like Framing", RFC 2687, September 1999.

[4] Casner, S. and V. Jacobson, "Compressing IP/UDP/RTP Headers for Low-Speed Serial Links", RFC 2508, February 1999.

[5] Wroclawski, J., "Specification of the Controlled-Load Network Element Service", RFC 2211, September 1997.

[6] Partridge, C. and R. Guerin, "Specification of Guaranteed Quality of Service", RFC 2212, September 1997.

[7] Shenker, S. and J. Wroclawski, "General Characterization Parameters for Integrated Service Network Elements", RFC 2215, September 1997.

[8] Jacobson, V., "TCP/IP Compression for Low-Speed Serial Links", RFC 1144, February 1990.

[9] B. Davie et al. "Integrated Services in the Presence of Compressible Flows", Work in Progress (draft-davie-intserv-compress-00.txt), Feb. 1999.

[10] Engan, M., Casner, S. and C. Bormann, "IP Header Compression over PPP", RFC 2509, February 1999.

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

[12] Sklower, K., Lloyd, B., McGregor, G., Carr, D. and T. Coradettim, "The PPP Multilink Protocol (MP)", RFC 1990, August 1996.

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

[14] Braden, R. and L. Zhang, "Resource ReSerVation Protocol (RSVP) -- Version 1 Message Processing Rules", RFC 2209, September 1997.

7. Authors' Addresses

Steve Jackowski
Deterministic Networks, Inc.
245M Mt Hermon Rd, #140
Scotts Valley, CA 95060
USA

Phone: +1 (408) 813 6294
EMail: stevej@DeterministicNetworks.com

David Putzolu
Intel Architecture Labs (IAL)
JF3-206-H10
2111 NE 25th Avenue
Hillsboro, OR 97124-5961
USA

Phone: +1 (503) 264 4510
EMail: David.Putzolu@intel.com

Eric S. Crawley
Argon Networks, Inc.
25 Porter Road
Littleton, MA 01460
USA

Phone: +1 (978) 486-0665
EMail: esc@argon.com

Bruce Davie
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824
USA

Phone: +1 (978) 244 8921
EMail: bdavie@cisco.com

Acknowledgements

This document draws heavily on the work of the ISSLL WG of the IETF.

Appendix A. Admission Control Considerations for POTS Modems

The protocols used in current implementations of POTS modems can exhibit significant changes in link rate and delay over the duration of a connection. Admission control and link scheduling algorithms used with these devices MUST be prepared to compensate for this variability in order to provide a robust implementation of integrated services.

Link rate on POTS modems is typically reported at connection time. This value may change over the duration of the connection. The v.34 protocol, used in most POTS modems, is adaptive to link conditions, and is able to recalibrate transmission rate multiple times over the duration of a connection. Typically this will result in a small (˜10%) increase in transmission rate over the initial connection within the first minute of a call. It is important to note, however, that other results are possible as well, including decreases in available bandwidth. Admission control algorithms MUST take such changes into consideration as they occur, and implementations MUST be able to gracefully handle the pathological case where link rate actually drops below the currently reserved capacity of a link.

Delay experienced by traffic over POTS modems can vary significantly over time. Unlike link rate, the delay often does not converge to a stable value. The v.42 protocol is used in most POTS modems to provide link-layer reliability. This reliability, which is implemented via retransmission, can cause frames to experience significant delays. Retransmissions also implicitly steal link bandwidth from other traffic. These delays and reductions in link bandwidth make it extremely difficult to honor a guaranteed service reservation. On a link that is actually lightly or moderately loaded, a controlled load service can to some extent accept such events as part of the behavior of a lightly loaded link. Unfortunately, as actual link utilization increases, v.42 retransmissions have the potential of stealing larger and larger fractions of available link bandwidth; making even controlled load service difficult to offer at high link utilization when retransmissions occur.

9. Full Copyright Statement

Acknowledgement

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