Internet Engineering Task Force   Ed Ellesson
INTERNET DRAFT   Steven Blake
draft-ellesson-tos-00.txt   IBM
                                                  November 1997
                                                  Expires May 1998

A Proposal for the Format and Semantics of the TOS Byte and Traffic Class Byte in IPv4 and IPv6 Headers

Status of this Memo

   This document is an Internet-Draft. 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."

   To view the entire list of current Internet-Drafts, please check the
   "1id-abstracts.txt" listing contained in the Internet-Drafts Shadow
   Directories on ftp.is.co.za (Africa), ftp.nordu.net (Europe),
   munnari.oz.au (Pacific Rim), ds.internic.net (US East Coast), or
   ftp.isi.edu (US West Coast).

Abstract

   This draft proposes an arrangement of fields in the IPv4 TOS byte
   [RFC795, RFC1349], and in the IPv6 Traffic Class byte [IPv6], and
   proposes a definition for their associated semantics. The intention
   is to enable the preservation of currently useful differential class-
   based queueing behavior on existing network devices, while
   simultaneously enabling new methods of bandwidth allocation and
   policing via drop preference, all within the context of flows which
   may be encrypted at the IP layer using IPSEC.

   (Note: the IPv6 Class field has recently been expanded to eight bits,
   but this in not yet available in any version of the specification).

1. Introduction

   Many intranets have deployed class-based queuing and/or class-based
   buffer management for a number of years. Routers and/or other
   network edge devices in these networks typically classify each packet
   received based on IP and TCP/UDP header fields which offer clues
   as to the application type or traffic type, and which therefore

reflect the relative business importance of the communicating
application. These clues include the source/destination port
numbers, protocol id, and source/destination addresses. These
classification criteria are then used to direct packets to
statically configured queues, or to identify the packets for drop
eligibility within the network device.

With the advent of IPSEC encryption of the IP payload, it is no
longer possible for routers and other network devices to look inside
the payload of an encrypted IP packet to extract the protocol id or
source/destination port numbers for the purpose of learning the
traffic type. As a result, the functionality of packet
classification and differential queueing/dropping based on traffic
type is lost when migrating applications to a network that supports
IPSEC, unless a different mechanism is enabled for packet
classification.

(The other circumstance when the classification information is lost
is IP fragmentation).

Commercial enterprises which are considering migrating their private
corporate network traffic to use the Internet, must ensure the
security and privacy of their traffic, while, at the same time,
preserving the precedence of certain traffic types over others in
order to support service level agreements with their users. If no
solution is standardized, either the deployment of VPN's and
extranets based on IPSEC will be impeded, or the current ability to
prioritize by traffic type will be lost.

2. Requirements

The following requirements drove the proposed solution:

-ability to prioritize packets by traffic class, for both delay and
 drop preference, in a standardized, multivendor-interoperable way.

-ability to support explicit congestion notification, as proposed by
 [ECN94, ECN97, Clark97].

-ability to isolate certain transport protocols (especially those
 using non-TCP congestion control algorithms) and/or traffic types
 into separate traffic classes for individualized queueing treatment
 in a network device, according to the needs of individual users and
 network operators and network administrators in the context of
 intranets, extranets, and VPN's [Blake97].

-ability to provision minimum bandwidth available to specific traffic
 classes (for starvation avoidance), either via marking at the
 network edge for dropping priority (new capability), or via setting
 dropping thresholds, leaky bucket parameters, and/or minimum drain
 rate in each queue that is dedicated to a specific traffic class

(existing capability extended to work within the IPSEC context).

   -ability to mark packets for differential service treatment at the
    packet source, or at an intervening network device, such as a router
    or traffic shaper/policer, for example, at the edge of the access to
    an ISP's wide area network.

   -avoid introduction of a migration penalty for current best effort
    users of a network which has partially deployed profile-meter
    marking.

   -ability to provide all the above functionality in the context of a
    network that supports IPSEC encryption. (Encryption may be
    implemented either at the packet stream source, or in a downstream
    device. Packet marking, via the TOS byte or the Traffic Class byte,
    must be introduced at, or upstream of, the encrypting device).

   -same semantics for IPv4 and IPv6, to enable coexistence/migration.

3. Proposed Solution

   IPv4 TOS Byte (and IPv6 Traffic Class Byte) Definition:

```
              0   1   2   3   4   5   6   7
            +---+---+---+---+---+---+---+---+
            |CE |ECT|DP | Service Class |MBZ|
            +---+---+---+---+---+---+---+---+
```

   Explicit Congestion Notification (ECN):

     CE: Congestion Experienced (1 bit)
         (router-, or traffic shaper/policer-settable)
        - 0: no congestion experienced
        - 1: congestion experienced

     ECT: ECN-capable transport (1 bit):
          (source host-settable)
        - 0: transport protocol not ECN-capable
        - 1: transport protocol ECN-capable

    Notes on Explicit Congestion Notification:

        -There are two proposals for the use of the CE bit. In [ECN94]
         and [ECN97], the CE bit is set stochastically based on random
         early detection of congestion (when ECT is set) [RED]. In
         [Clark97], the CE bit is set deterministically for every packet
         under impending congestion, and then the CE bit is filtered
         downstream by a receiver profile meter. The proposed mapping
         supports both approaches. Routers shall support the behavior
         specified in [ECN97] by default. Routers shall support the
         behavior specified in [Clark97] as a configuration option (to
         allow the deployment of receiver-based service allocation).

   Drop Preference (DP) (1 bit):

      - 0: not discard-eligible/in-profile/loss-sensitive
      - 1: discard eligible/out-of-profile/loss-insensitive

    Notes on Drop Preference:

      -[Clark97]- and [Feng97]-style service-allocation
       implementations should use the DP bit to signal in- or out-of-
       profile packets. Integrated Services policers shall use the DP
       bit to signal non-conformant packets. Routers shall not re-
       order packets within a flow when the DP bit is toggled.

   Explicit Service Class (4 bits):

      - 0001: Delay Insensitive (minimize cost/worse than best effort)
      - 0000: Normal   (best effort)
      - 1000: Interactive Delay (low average delay)
      - 1001: Low Maximum Delay (ie: low delay/jitter, eg. real time)
      - 0010: Network Control   (maximize reliability)
      - 0100: Maximize Throughput
      - 0011: Network-specific 1
      - 0101: Network-specific 2
      - 0110: Network-specific 3
      - 0111: Network-specific 4
      - 1010: Network-specific 5
      - 1011: IntServ (low priority),   Network-specific 6
      - 1100: IntServ (medium priority), Network-specific 7
      - 1101: IntServ (high priority),   Network-specific 8
      - 1110: Reserved 1
      - 1111: Reserved 2

    Notes on Explicit Service Class:

      -The specific value assignments noted above were chosen to
       preserve backwards compatibility with [RFC1349].

      -The value assignments noted above are shown out of numerical
       order to highlight proximity of interpretation.

      -The first four service classes noted above are ranked in order
       of increasing delay priority.

      -Any of the above service classes may have specific minimum
       bandwidth allocations, delay priorities, and/or drop
       thresholds configured within the routers. The means of
       configuring these parameters are beyond the scope of this
       specification.

       (One possible implementation method is to direct packets
       with a specific service class to an associated class-based
       queue [CBQ]. This permits the service class to act as an index

into a queue that has been configured for a particular traffic
type, in the same way that the port numbers are used today on
commercial intranets. Note that the combination of source/
destination address, protocol id, and service class can be
used in the same way, that is, as an index into a queue.)

-The network-specific code points are available to be used with
 any application/traffic type, or set of application types,
 agreed to by the network administrator/network operator and
 the end user/service subscriber. The network administrator
 may provision resources for each network-specific service class
 as appropriate to provide the level of performance required for
 traffic mapped to that class. The network administrator may
 associate any network-specific service class with a particular
 drop or delay priority. By specifying minimum bandwidths per-
 class (mechanism for doing this is outside the scope of this
 draft), the network administrator can avoid starvation of lower
 priority flows.

 -The network-specific code points are not ranked in any implied
  order of loss, delay, or throughput priority.

-Routing and other network control protocols using this mapping
 which require prioritized handling or reliable delivery by the
 network shall be marked with service class '0010' (Network
 Control).

-IntServ code values '1011', '1100', '1101', are intended for
 use by packets of a reserved Integrated Services flow where
 RSVP aggregation is deployed [GBH97]. Packets within an
 aggregated reservation should be mapped to one of the three
 service classes (depending on the IntServ traffic class) to
 facilitate packet classification. The router implementation
 must isolate traffic in these service classes from traffic
 which has not been policed at an RSVP aggregation point. The
 IntServ service class implementations must prevent non-
 conformant packets (marked by DP) from degrading the QoS of
 other flows within the same service class.

 If RSVP aggregation is not deployed, then these code values
 are available as network-specific service classes. If
 additional IntServ service classes are desired, they may be
 allocated from the network-specific code points at the
 discretion of the network administrator.

-Service class-specific routing (i.e., TOS routing) may be
 implemented at the option of the network administrator.
 Specification of the routing metrics to be associated with
 each service class is beyond the scope of this draft.

   -Interaction of network-specific code points with DP and ECN
    field values:

        If supported by network policy, an edge device may instead
        use the DP marking of out-of-profile traffic to provide
        minimum bandwidth guarantees, rather than using minimum
        bandwidths configured in each router.

        In addition, drop and congestion control may be provided
        individually within each service class. That is, packets may
        be marked for drop eligibility or for explicit congestion
        notification within a specific allocated bandwidth,
        configured for a specific service class of traffic, rather
        than relative to the entire bandwidth available on an
        outgoing interface.

        The service class implementation may choose to ignore the CE,
        ECT, or DP bits.

   -Reserved codepoints are available for future standardization or
    experimentation.

MBZ (1bit):

    Must be zero. Reserved for use on experimental networks with
    TOS Byte or Traffic Class Byte definition other than above.

4. Network Interoperability

Service providers which exchange traffic and support differentiated
services via service-class-value-marked packets, should either agree
to compatible definitions for Network-specific values, or they should
agree to map the Network-specific values into one of the standardized
values at their interconnection point.

A scaleable administrative mechanism for managing the mapping of
traffic type to service class, and from service class to service
class (across a domain boundary) is key to the manageable deployment
of this solution on a wide scale. Scaleable administrative
mechanisms are beyond the scope of this draft.

5. Backwards Compatibility With RFC 795

[Ferg97] suggests the usage of the IPv4 Precedence field to signify
the drop preference of in-profile or out-of-profile packets, as
defined for example by [Clark97]. Out-of-profile packets would be
marked with lower precedence than in-profile packets. Routers which
implemented preferential discard based on the semantics of [RFC795]
would preferentially discard out-of-profile packets in times of
impending congestion.

A potential problem that may occur during the phase of partial
deployment of traffic profile meters is that the bulk of existing
best-effort traffic is marked with the "Routine" precedence value
'000'. This un-metered traffic which enters a network implementing
precedence dropping would be treated as out-of-profile. It is not
clear that this is always the correct choice, which motivated our
decision to abandon the existing semantics of the Precedence field
and explicitly allocate a drop preference bit.

However, it is also the case that many routing protocol
implementations transmit their packets with "Internetwork Control"
precedence '110', as specified in [RFC1812]. During a transition
period where not all routers have been upgraded to use the proposed
service class mapping for network control (service class '0010'), it
may be valuable to provide backwards compatibility with RFC 795
Precedence semantics.

We know of two approaches to achieve this:

-utilize the MBZ bit as an indicator of the version of the TOS/Class
 mapping semantics

   - 0: compliant with [RFC795] (Precedence) and RFC 1349 (TOS)
   - 1: compliant with the proposed mapping

-utilize the service class '0000' to indicate that the mapping
 is compliant with [RFC795] (Precedence) and RFC 1349 (TOS).

In the first approach, the MBZ bit is consumed as a version
indicator.

In the second approach, the service class '0000' is consumed, but the
MBZ bit remains reserved for future specification.

The authors believe it to be the case that existing routing protocols
typically use a zero TOS value, and further, that most best-
effort traffic utilizes the zero TOS value. Existing routing traffic
transmitted with a non-zero precedence and zero TOS would continue to
receive preferential queueing by routers which implemented either
of the above approaches.

Best-effort traffic which was transmitted with zero precedence and
zero TOS (which we believe to include the bulk of Internet data
traffic) and which was not metered would not receive degraded service
from routers which implemented either of the above TOS approaches
(this could be configurable in routers implementing the proposed
mapping).

Traffic generated by hosts or routers which have not implemented the
proposed TOS semantics and which utilize a non-zero TOS value would
be mapped into the corresponding service class by routers
implementing the proposed semantics (unless the traffic was remapped

upstream by some other device). Since the proposed semantics are

compatible with the TOS classes defined in [RFC1349], this is no
more of a potential problem than in the case where hosts or routers
which have implemented the proposed TOS mapping are able to send
traffic mapped into a service class without network authorization or
monitoring. (A scaleable mechanism for network devices to remotely
acquire network authorization policy are beyond the scope of this
draft.)

One effect of this change is that the "Normal" service class can no
longer be utilized by hosts, routers, and profile-meters implementing
the new proposed TOS semantics (since "Normal" equals "old"
semantics). Best-effort traffic which does not require service
differentiation, but wishes to take advantage of ECN, for example,
would need to specify an alternative service class, such as
"Delay Insensitive" ('0001'), or one of the reserved classes (one of
these options would be standardized).

It should be noted that backwards compatibility is proposed *FOR IPV4
ONLY*. The [RFC795] Precedence semantics would never be utilized by
IPv6 routers, and the "Normal" service class ('0000') would be
available for best-effort traffic not requiring service
differentiation.

6. Security Considerations

Security considerations are not discussed in this memo.

7. References

[Blake97] S. Blake, "Some Issues and Applications of Packet Marking
           for Differentiated Services", Internet Draft
           <draft-blake-diffserv-marking-00.txt>, November 1997.

[Clark97] D. Clark and J. Wroclawski, "An Approach to Service
           Allocation in the Internet", Internet Draft
           <draft-clark-diff-svc-alloc-00.txt>, July 1997.

[CBQ]   S. Floyd and V. Jacobson, "Link-sharing and Resource
           Management Models for Packet Networks", IEEE/ACM
           Transactions on Networking, Vol. 3 no. 4, pp. 365-386,
           August 1995.

[ECN94]   S. Floyd, "TCP and Explicit Congestion Notification",
           ACM Computer Communications Review, Vol. 24 no. 5, pp. 10-
           23, October 1994.

[ECN97]   K. Ramakrishnan and S. Floyd, "A Proposal to Add Explicit
           Congestion Notification (ECN) to IPv6 and to TCP",
           Internet Draft <draft-kksjf-ecn-00.txt>, November 1997.

[Feng97]   W. Feng, D. Kandlur, D. Saha, and K. Shin, "Adaptive
            Packet Marking for Providing Differentiated Services in
            the Internet", Univ. Michigan Technical Report
            CSE-TR-347-97, October 1997,
            http://www.eecs.umich.edu/~wuchang/work/pmg.ps.Z.

[Ferg97]   P. Ferguson, "Simple Differential Services: IP TOS and
            Precedence, Delay Indication, and Drop Preference,
            Internet Draft <draft-ferguson-delay-drop-00.txt>,
            November 1997.

[GBH97]   R. Guerin, S. Blake, and S. Herzog, "Aggregating RSVP-
            based QoS Requests", Internet Draft
            <draft-guerin-aggreg-rsvp-00.txt>, November 1997.

[IPv6]   S. Deering and R. Hinden, "Internet Protocol, Version 6
            (IPv6) Specification", Internet Draft
            <draft-ietf-ipngwg-ipv6-spec-v2-00.txt>, July 1997.

[RED]   S. Floyd and V. Jacobson, "Random Early Detection Gateways
            for Congestion Avoidance", IEEE/ACM Transactions on
            Networking, August 1993.

[RFC795]   J. Postel, "Service Mappings", Internet RFC 795, September
            1981.

[RFC1349]   P. Almquist, "Type of Service in the Internet Protocol
            Suite", Internet RFC 1349, July 1992.

[RFC1812]   F. Baker, editor, "Requirements for IP Version 4 Routers",
            Internet RFC 1812, June 1995.

[SIMA]   K. Kilkki, "Simple Integrated Media Access (SIMA)",
            Internet Draft <draft-kalevi-simple-media-access-01.txt>,
            June 1997.

Appendix A: Specifying Multiple Drop Preference Levels

   Because the concept of service class is completely general, it is
   possible to utilize different service classes to represent different
   drop preference levels, as would be needed for example by [SIMA].
   Shown below is a possible mapping implementing eight separate drop
   preference levels (where higher drop preference results in higher
   probability of loss):

DP	Service Class (Value)	(Name)	Drop Preference
0	0011	Network-specific 1	0
1	0011	Network-specific 1	1
0	0101	Network-specific 2	2
1	0101	Network-specific 2	3
0	0110	Network-specific 3	4
1	0110	Network-specific 3	5
0	0111	Network-specific 4	6
1	0111	Network-specific 4	7

Appendix B: Specifying Multiple Delay Priorities

   As was mentioned in Sec. 3, there are four service classes defined
   which specify relative delay priority (three for IPv4 if backwards
   compatibility with [RFC795] is required):

        - 0001: Delay Insensitive (minimize cost/worse than best effort)
        - 0000: Normal   (best effort)
        - 1000: Interactive Delay (low average delay)
        - 1001: Low Maximum Delay (ie: low delay/jitter, eg. real time)

   If additional provisioned levels of delay priority are required, they
   can be implemented using the Network-specific service classes. Shown
   below is a possible mapping implementing eight separate delay
   priorities (where a higher priority results in lower average/maximum
   delay):

Service Class (Value)	(Name)	Delay Priority
0001	Delay Insensitive	0
0011	Network-specific 1	1
0101	Network-specific 2	2
1000	Interactive Delay	3
0110	Network-specific 3	4
0111	Network-specific 4	5
1001	Low Maximum Delay	6
1010	Network-specific 5	7

Authors' Addresses

    Ed Ellesson
    JDGA/501
    IBM Corporation
    4205 S. Miami Blvd.
    Research Triangle Park, NC 27709
    Phone: +1-919-254-4115
    Fax: +1-919-254-6243
    E-mail: ellesson@raleigh.ibm.com

    Steven Blake
    E95/664
    IBM Corporation
    800 Park Offices Drive
    Research Triangle Park, NC 27709
    Phone: +1-919-254-2030
    Fax: +1-919-254-5483
    E-mail: slblake@raleigh.ibm.com