Network Working Group
Request for Comments: 2475
Category: Informational

S. Blake
Torrent Networking Technologies
D. Black
EMC Corporation
M. Carlson
Sun Microsystems
E. Davies
Nortel UK
Z. Wang
Bell Labs Lucent Technologies
W. Weiss
Lucent Technologies

December 1998

An Architecture for Differentiated Services

Status of this Memo

Copyright Notice

Abstract

This document defines an architecture for implementing scalable service differentiation in the Internet. This architecture achieves scalability by aggregating traffic classification state which is conveyed by means of IP-layer packet marking using the DS field [DSFIELD]. Packets are classified and marked to receive a particular per-hop forwarding behavior on nodes along their path. Sophisticated classification, marking, policing, and shaping operations need only be implemented at network boundaries or hosts. Network resources are allocated to traffic streams by service provisioning policies which govern how traffic is marked and conditioned upon entry to a differentiated services-capable network, and how that traffic is forwarded within that network. A wide variety of services can be implemented on top of these building blocks.

Table of Contents

1. Introduction

1.1 Overview

This document defines an architecture for implementing scalable service differentiation in the Internet. A "Service" defines some significant characteristics of packet transmission in one direction across a set of one or more paths within a network. These

characteristics may be specified in quantitative or statistical terms of throughput, delay, jitter, and/or loss, or may otherwise be specified in terms of some relative priority of access to network resources. Service differentiation is desired to accommodate heterogeneous application requirements and user expectations, and to permit differentiated pricing of Internet service.

This architecture is composed of a number of functional elements implemented in network nodes, including a small set of per-hop forwarding behaviors, packet classification functions, and traffic conditioning functions including metering, marking, shaping, and policing. This architecture achieves scalability by implementing complex classification and conditioning functions only at network boundary nodes, and by applying per-hop behaviors to aggregates of traffic which have been appropriately marked using the DS field in the IPv4 or IPv6 headers [DSFIELD]. Per-hop behaviors are defined to permit a reasonably granular means of allocating buffer and bandwidth resources at each node among competing traffic streams. Per-application flow or per-customer forwarding state need not be maintained within the core of the network. A distinction is maintained between:

- the service provided to a traffic aggregate,
- the conditioning functions and per-hop behaviors used to realize services,
- the DS field value (DS codepoint) used to mark packets to select a per-hop behavior, and
- the particular node implementation mechanisms which realize a per-hop behavior.

Service provisioning and traffic conditioning policies are sufficiently decoupled from the forwarding behaviors within the network interior to permit implementation of a wide variety of service behaviors, with room for future expansion.

This architecture only provides service differentiation in one direction of traffic flow and is therefore asymmetric. Development of a complementary symmetric architecture is a topic of current research but is outside the scope of this document; see for example [EXPLICIT].

Sect. 1.2 is a glossary of terms used within this document. Sec. 1.3 lists requirements addressed by this architecture, and Sec. 1.4 provides a brief comparison to other approaches for service differentiation. Sec. 2 discusses the components of the architecture

in detail. Sec. 3 proposes guidelines for per-hop behavior specifications. Sec. 4 discusses interoperability issues with nodes and networks which do not implement differentiated services as defined in this document and in [DSFIELD]. Sec. 5 discusses issues with multicast service delivery. Sec. 6 addresses security and tunnel considerations.

1.2 Terminology

This section gives a general conceptual overview of the terms used in this document. Some of these terms are more precisely defined in later sections of this document.

Behavior Aggregate (BA)	a DS behavior aggregate.
BA classifier	a classifier that selects packets based only on the contents of the DS field.
Boundary link	a link connecting the edge nodes of two domains.
Classifier	an entity which selects packets based on the content of packet headers according to defined rules.
DS behavior aggregate	a collection of packets with the same DS codepoint crossing a link in a particular direction.
DS boundary node	a DS node that connects one DS domain to a node either in another DS domain or in a domain that is not DS-capable.
DS-capable	capable of implementing differentiated services as described in this architecture; usually used in reference to a domain consisting of DS-compliant nodes.
DS codepoint	a specific value of the DSCP portion of the DS field, used to select a PHB.
DS-compliant	enabled to support differentiated services functions and behaviors as defined in [DSFIELD], this document, and other differentiated services documents; usually used in reference to a node or device.

DS domain	a DS-capable domain; a contiguous set of nodes which operate with a common set of service provisioning policies and PHB definitions.
DS egress node	a DS boundary node in its role in handling traffic as it leaves a DS domain.
DS ingress node	a DS boundary node in its role in handling traffic as it enters a DS domain.
DS interior node	a DS node that is not a DS boundary node.
DS field	the IPv4 header TOS octet or the IPv6 Traffic Class octet when interpreted in conformance with the definition given in [DSFIELD]. The bits of the DSCP field encode the DS codepoint, while the remaining bits are currently unused.
DS node	a DS-compliant node.
DS region	a set of contiguous DS domains which can offer differentiated services over paths across those DS domains.
Downstream DS domain	the DS domain downstream of traffic flow on a boundary link.
Dropper	a device that performs dropping.
Dropping	the process of discarding packets based on specified rules; policing.
Legacy node	a node which implements IPv4 Precedence as defined in [RFC791,RFC1812] but which is otherwise not DS-compliant.
Marker	a device that performs marking.
Marking	the process of setting the DS codepoint in a packet based on defined rules; pre-marking, re-marking.
Mechanism	a specific algorithm or operation (e.g., queueing discipline) that is implemented in a node to realize a set of one or more per-hop behaviors.

Meter	a device that performs metering.
Metering	the process of measuring the temporal properties (e.g., rate) of a traffic stream selected by a classifier. The instantaneous state of this process may be used to affect the operation of a marker, shaper, or dropper, and/or may be used for accounting and measurement purposes.
Microflow	a single instance of an application-to-application flow of packets which is identified by source address, source port, destination address, destination port and protocol id.
MF Classifier	a multi-field (MF) classifier which selects packets based on the content of some arbitrary number of header fields; typically some combination of source address, destination address, DS field, protocol ID, source port and destination port.
Per-Hop-Behavior (PHB)	the externally observable forwarding behavior applied at a DS-compliant node to a DS behavior aggregate.
PHB group	a set of one or more PHBs that can only be meaningfully specified and implemented simultaneously, due to a common constraint applying to all PHBs in the set such as a queue servicing or queue management policy. A PHB group provides a service building block that allows a set of related forwarding behaviors to be specified together (e.g., four dropping priorities). A single PHB is a special case of a PHB group.
Policing	the process of discarding packets (by a dropper) within a traffic stream in accordance with the state of a corresponding meter enforcing a traffic profile.
Pre-mark	to set the DS codepoint of a packet prior to entry into a downstream DS domain.

Provider DS domain	the DS-capable provider of services to a source domain.
Re-mark	to change the DS codepoint of a packet, usually performed by a marker in accordance with a TCA.
Service	the overall treatment of a defined subset of a customer's traffic within a DS domain or end-to-end.
Service Level Agreement (SLA)	a service contract between a customer and a service provider that specifies the forwarding service a customer should receive. A customer may be a user organization (source domain) or another DS domain (upstream domain). A SLA may include traffic conditioning rules which constitute a TCA in whole or in part.
Service Provisioning Policy	a policy which defines how traffic conditioners are configured on DS boundary nodes and how traffic streams are mapped to DS behavior aggregates to achieve a range of services.
Shaper	a device that performs shaping.
Shaping	the process of delaying packets within a traffic stream to cause it to conform to some defined traffic profile.
Source domain	a domain which contains the node(s) originating the traffic receiving a particular service.
Traffic conditioner	an entity which performs traffic conditioning functions and which may contain meters, markers, droppers, and shapers. Traffic conditioners are typically deployed in DS boundary nodes only. A traffic conditioner may re-mark a traffic stream or may discard or shape packets to alter the temporal characteristics of the stream and bring it into compliance with a traffic profile.

Traffic conditioning	control functions performed to enforce rules specified in a TCA, including metering, marking, shaping, and policing.
Traffic Conditioning Agreement (TCA)	an agreement specifying classifier rules and any corresponding traffic profiles and metering, marking, discarding and/or shaping rules which are to apply to the traffic streams selected by the classifier. A TCA encompasses all of the traffic conditioning rules explicitly specified within a SLA along with all of the rules implicit from the relevant service requirements and/or from a DS domain's service provisioning policy.
Traffic profile	a description of the temporal properties of a traffic stream such as rate and burst size.
Traffic stream	an administratively significant set of one or more microflows which traverse a path segment. A traffic stream may consist of the set of active microflows which are selected by a particular classifier.
Upstream DS domain	the DS domain upstream of traffic flow on a boundary link.

1.3 Requirements

The history of the Internet has been one of continuous growth in the number of hosts, the number and variety of applications, and the capacity of the network infrastructure, and this growth is expected to continue for the foreseeable future. A scalable architecture for service differentiation must be able to accommodate this continued growth.

The following requirements were identified and are addressed in this architecture:

- should accommodate a wide variety of services and provisioning policies, extending end-to-end or within a particular (set of) network(s),
- should allow decoupling of the service from the particular application in use,

- should work with existing applications without the need for application programming interface changes or host software modifications (assuming suitable deployment of classifiers, markers, and other traffic conditioning functions),
- should decouple traffic conditioning and service provisioning functions from forwarding behaviors implemented within the core network nodes,
- should not depend on hop-by-hop application signaling,
- should require only a small set of forwarding behaviors whose implementation complexity does not dominate the cost of a network device, and which will not introduce bottlenecks for future high-speed system implementations,
- should avoid per-microflow or per-customer state within core network nodes,
- should utilize only aggregated classification state within the network core,
- should permit simple packet classification implementations in core network nodes (BA classifier),
- should permit reasonable interoperability with non-DS-compliant network nodes,
- should accommodate incremental deployment.

1.4 Comparisons with Other Approaches

The differentiated services architecture specified in this document can be contrasted with other existing models of service differentiation. We classify these alternative models into the following categories: relative priority marking, service marking, label switching, Integrated Services/RSVP, and static per-hop classification.

Examples of the relative priority marking model include IPv4 Precedence marking as defined in [RFC791], 802.5 Token Ring priority [TR], and the default interpretation of 802.1p traffic classes [802.1p]. In this model the application, host, or proxy node selects a relative priority or "precedence" for a packet (e.g., delay or discard priority), and the network nodes along the transit path apply the appropriate priority forwarding behavior corresponding to the priority value within the packet's header. Our architecture can be considered as a refinement to this model, since we more clearly

specify the role and importance of boundary nodes and traffic conditioners, and since our per-hop behavior model permits more general forwarding behaviors than relative delay or discard priority.

An example of a service marking model is IPv4 TOS as defined in [RFC1349]. In this example each packet is marked with a request for a "type of service", which may include "minimize delay", "maximize throughput", "maximize reliability", or "minimize cost". Network nodes may select routing paths or forwarding behaviors which are suitably engineered to satisfy the service request. This model is subtly different from our architecture. Note that we do not describe the use of the DS field as an input to route selection. The TOS markings defined in [RFC1349] are very generic and do not span the range of possible service semantics. Furthermore, the service request is associated with each individual packet, whereas some service semantics may depend on the aggregate forwarding behavior of a sequence of packets. The service marking model does not easily accommodate growth in the number and range of future services (since the codepoint space is small) and involves configuration of the "TOS->forwarding behavior" association in each core network node. Standardizing service markings implies standardizing service offerings, which is outside the scope of the IETF. Note that provisions are made in the allocation of the DS codepoint space to allow for locally significant codepoints which may be used by a provider to support service marking semantics [DSFIELD].

Examples of the label switching (or virtual circuit) model include Frame Relay, ATM, and MPLS [FRELAY, ATM]. In this model path forwarding state and traffic management or QoS state is established for traffic streams on each hop along a network path. Traffic aggregates of varying granularity are associated with a label switched path at an ingress node, and packets/cells within each label switched path are marked with a forwarding label that is used to lookup the next-hop node, the per-hop forwarding behavior, and the replacement label at each hop. This model permits finer granularity resource allocation to traffic streams, since label values are not globally significant but are only significant on a single link; therefore resources can be reserved for the aggregate of packets/ cells received on a link with a particular label, and the label switching semantics govern the next-hop selection, allowing a traffic stream to follow a specially engineered path through the network. This improved granularity comes at the cost of additional management and configuration requirements to establish and maintain the label switched paths. In addition, the amount of forwarding state maintained at each node scales in proportion to the number of edge nodes of the network in the best case (assuming multipoint-to-point

label switched paths), and it scales in proportion with the square of the number of edge nodes in the worst case, when edge-edge label switched paths with provisioned resources are employed.

The Integrated Services/RSVP model relies upon traditional datagram forwarding in the default case, but allows sources and receivers to exchange signaling messages which establish additional packet classification and forwarding state on each node along the path between them [RFC1633, RSVP]. In the absence of state aggregation, the amount of state on each node scales in proportion to the number of concurrent reservations, which can be potentially large on high-speed links. This model also requires application support for the RSVP signaling protocol. Differentiated services mechanisms can be utilized to aggregate Integrated Services/RSVP state in the core of the network [Bernet].

A variant of the Integrated Services/RSVP model eliminates the requirement for hop-by-hop signaling by utilizing only "static" classification and forwarding policies which are implemented in each node along a network path. These policies are updated on administrative timescales and not in response to the instantaneous mix of microflows active in the network. The state requirements for this variant are potentially worse than those encountered when RSVP is used, especially in backbone nodes, since the number of static policies that might be applicable at a node over time may be larger than the number of active sender-receiver sessions that might have installed reservation state on a node. Although the support of large numbers of classifier rules and forwarding policies may be computationally feasible, the management burden associated with installing and maintaining these rules on each node within a backbone network which might be traversed by a traffic stream is substantial.

Although we contrast our architecture with these alternative models of service differentiation, it should be noted that links and nodes employing these techniques may be utilized to extend differentiated services behaviors and semantics across a layer-2 switched infrastructure (e.g., 802.1p LANs, Frame Relay/ATM backbones) interconnecting DS nodes, and in the case of MPLS may be used as an alternative intra-domain implementation technology. The constraints imposed by the use of a specific link-layer technology in particular regions of a DS domain (or in a network providing access to DS domains) may imply the differentiation of traffic on a coarser grain basis. Depending on the mapping of PHBs to different link-layer services and the way in which packets are scheduled over a restricted set of priority classes (or virtual circuits of different category and capacity), all or a subset of the PHBs in use may be supportable (or may be indistinguishable).

2. Differentiated Services Architectural Model

The differentiated services architecture is based on a simple model where traffic entering a network is classified and possibly conditioned at the boundaries of the network, and assigned to different behavior aggregates. Each behavior aggregate is identified by a single DS codepoint. Within the core of the network, packets are forwarded according to the per-hop behavior associated with the DS codepoint. In this section, we discuss the key components within a differentiated services region, traffic classification and conditioning functions, and how differentiated services are achieved through the combination of traffic conditioning and PHB-based forwarding.

2.1 Differentiated Services Domain

A DS domain is a contiguous set of DS nodes which operate with a common service provisioning policy and set of PHB groups implemented on each node. A DS domain has a well-defined boundary consisting of DS boundary nodes which classify and possibly condition ingress traffic to ensure that packets which transit the domain are appropriately marked to select a PHB from one of the PHB groups supported within the domain. Nodes within the DS domain select the forwarding behavior for packets based on their DS codepoint, mapping that value to one of the supported PHBs using either the recommended codepoint->PHB mapping or a locally customized mapping [DSFIELD]. Inclusion of non-DS-compliant nodes within a DS domain may result in unpredictable performance and may impede the ability to satisfy service level agreements (SLAs).

A DS domain normally consists of one or more networks under the same administration; for example, an organization’s intranet or an ISP. The administration of the domain is responsible for ensuring that adequate resources are provisioned and/or reserved to support the SLAs offered by the domain.

2.1.1 DS Boundary Nodes and Interior Nodes

A DS domain consists of DS boundary nodes and DS interior nodes. DS boundary nodes interconnect the DS domain to other DS or non-DS-capable domains, whilst DS interior nodes only connect to other DS interior or boundary nodes within the same DS domain.

Both DS boundary nodes and interior nodes must be able to apply the appropriate PHB to packets based on the DS codepoint; otherwise unpredictable behavior may result. In addition, DS boundary nodes may be required to perform traffic conditioning functions as defined by a traffic conditioning agreement (TCA) between their DS domain and

the peering domain which they connect to (see Sec. 2.3.3).

Interior nodes may be able to perform limited traffic conditioning functions such as DS codepoint re-marking. Interior nodes which implement more complex classification and traffic conditioning functions are analogous to DS boundary nodes (see Sec. 2.3.4.4).

A host in a network containing a DS domain may act as a DS boundary node for traffic from applications running on that host; we therefore say that the host is within the DS domain. If a host does not act as a boundary node, then the DS node topologically closest to that host acts as the DS boundary node for that host's traffic.

2.1.2 DS Ingress Node and Egress Node

DS boundary nodes act both as a DS ingress node and as a DS egress node for different directions of traffic. Traffic enters a DS domain at a DS ingress node and leaves a DS domain at a DS egress node. A DS ingress node is responsible for ensuring that the traffic entering the DS domain conforms to any TCA between it and the other domain to which the ingress node is connected. A DS egress node may perform traffic conditioning functions on traffic forwarded to a directly connected peering domain, depending on the details of the TCA between the two domains. Note that a DS boundary node may act as a DS interior node for some set of interfaces.

2.2 Differentiated Services Region

A differentiated services region (DS Region) is a set of one or more contiguous DS domains. DS regions are capable of supporting differentiated services along paths which span the domains within the region.

The DS domains in a DS region may support different PHB groups internally and different codepoint->PHB mappings. However, to permit services which span across the domains, the peering DS domains must each establish a peering SLA which defines (either explicitly or implicitly) a TCA which specifies how transit traffic from one DS domain to another is conditioned at the boundary between the two DS domains.

It is possible that several DS domains within a DS region may adopt a common service provisioning policy and may support a common set of PHB groups and codepoint mappings, thus eliminating the need for traffic conditioning between those DS domains.

2.3 Traffic Classification and Conditioning

Differentiated services are extended across a DS domain boundary by establishing a SLA between an upstream network and a downstream DS domain. The SLA may specify packet classification and re-marking rules and may also specify traffic profiles and actions to traffic streams which are in- or out-of-profile (see Sec. 2.3.2). The TCA between the domains is derived (explicitly or implicitly) from this SLA.

The packet classification policy identifies the subset of traffic which may receive a differentiated service by being conditioned and/or mapped to one or more behavior aggregates (by DS codepoint re-marking) within the DS domain.

Traffic conditioning performs metering, shaping, policing and/or re-marking to ensure that the traffic entering the DS domain conforms to the rules specified in the TCA, in accordance with the domain's service provisioning policy. The extent of traffic conditioning required is dependent on the specifics of the service offering, and may range from simple codepoint re-marking to complex policing and shaping operations. The details of traffic conditioning policies which are negotiated between networks is outside the scope of this document.

2.3.1 Classifiers

Packet classifiers select packets in a traffic stream based on the content of some portion of the packet header. We define two types of classifiers. The BA (Behavior Aggregate) Classifier classifies packets based on the DS codepoint only. The MF (Multi-Field) classifier selects packets based on the value of a combination of one or more header fields, such as source address, destination address, DS field, protocol ID, source port and destination port numbers, and other information such as incoming interface.

Classifiers are used to "steer" packets matching some specified rule to an element of a traffic conditioner for further processing. Classifiers must be configured by some management procedure in accordance with the appropriate TCA.

The classifier should authenticate the information which it uses to classify the packet (see Sec. 6).

Note that in the event of upstream packet fragmentation, MF classifiers which examine the contents of transport-layer header fields may incorrectly classify packet fragments subsequent to the first. A possible solution to this problem is to maintain

fragmentation state; however, this is not a general solution due to the possibility of upstream fragment re-ordering or divergent routing paths. The policy to apply to packet fragments is outside the scope of this document.

2.3.2 Traffic Profiles

A traffic profile specifies the temporal properties of a traffic stream selected by a classifier. It provides rules for determining whether a particular packet is in-profile or out-of-profile. For example, a profile based on a token bucket may look like:

```
codepoint=X, use token-bucket r, b
```

The above profile indicates that all packets marked with DS codepoint X should be measured against a token bucket meter with rate r and burst size b. In this example out-of-profile packets are those packets in the traffic stream which arrive when insufficient tokens are available in the bucket. The concept of in- and out-of-profile can be extended to more than two levels, e.g., multiple levels of conformance with a profile may be defined and enforced.

Different conditioning actions may be applied to the in-profile packets and out-of-profile packets, or different accounting actions may be triggered. In-profile packets may be allowed to enter the DS domain without further conditioning; or, alternatively, their DS codepoint may be changed. The latter happens when the DS codepoint is set to a non-Default value for the first time [DSFIELD], or when the packets enter a DS domain that uses a different PHB group or codepoint->PHB mapping policy for this traffic stream. Out-of-profile packets may be queued until they are in-profile (shaped), discarded (policed), marked with a new codepoint (re-marked), or forwarded unchanged while triggering some accounting procedure. Out-of-profile packets may be mapped to one or more behavior aggregates that are "inferior" in some dimension of forwarding performance to the BA into which in-profile packets are mapped.

Note that a traffic profile is an optional component of a TCA and its use is dependent on the specifics of the service offering and the domain's service provisioning policy.

2.3.3 Traffic Conditioners

A traffic conditioner may contain the following elements: meter, marker, shaper, and dropper. A traffic stream is selected by a classifier, which steers the packets to a logical instance of a traffic conditioner. A meter is used (where appropriate) to measure the traffic stream against a traffic profile. The state of the meter

with respect to a particular packet (e.g., whether it is in- or out-of-profile) may be used to affect a marking, dropping, or shaping action.

When packets exit the traffic conditioner of a DS boundary node the DS codepoint of each packet must be set to an appropriate value.

Fig. 1 shows the block diagram of a classifier and traffic conditioner. Note that a traffic conditioner may not necessarily contain all four elements. For example, in the case where no traffic profile is in effect, packets may only pass through a classifier and a marker.

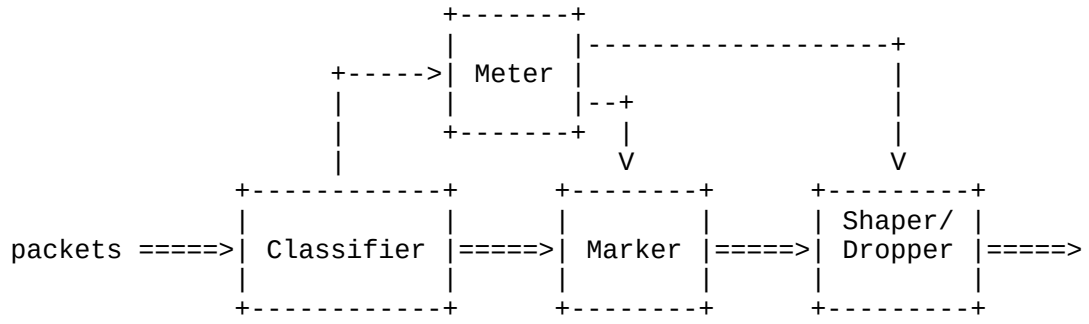

Fig. 1: Logical View of a Packet Classifier and Traffic Conditioner

2.3.3.1 Meters

Traffic meters measure the temporal properties of the stream of packets selected by a classifier against a traffic profile specified in a TCA. A meter passes state information to other conditioning functions to trigger a particular action for each packet which is either in- or out-of-profile (to some extent).

2.3.3.2 Markers

Packet markers set the DS field of a packet to a particular codepoint, adding the marked packet to a particular DS behavior aggregate. The marker may be configured to mark all packets which are steered to it to a single codepoint, or may be configured to mark a packet to one of a set of codepoints used to select a PHB in a PHB group, according to the state of a meter. When the marker changes the codepoint in a packet it is said to have "re-marked" the packet.

2.3.3.3 Shapers

Shapers delay some or all of the packets in a traffic stream in order to bring the stream into compliance with a traffic profile. A shaper usually has a finite-size buffer, and packets may be discarded if there is not sufficient buffer space to hold the delayed packets.

2.3.3.4 Droppers

Droppers discard some or all of the packets in a traffic stream in order to bring the stream into compliance with a traffic profile. This process is know as "policing" the stream. Note that a dropper can be implemented as a special case of a shaper by setting the shaper buffer size to zero (or a few) packets.

2.3.4 Location of Traffic Conditioners and MF Classifiers

Traffic conditioners are usually located within DS ingress and egress boundary nodes, but may also be located in nodes within the interior of a DS domain, or within a non-DS-capable domain.

2.3.4.1 Within the Source Domain

We define the source domain as the domain containing the node(s) which originate the traffic receiving a particular service. Traffic sources and intermediate nodes within a source domain may perform traffic classification and conditioning functions. The traffic originating from the source domain across a boundary may be marked by the traffic sources directly or by intermediate nodes before leaving the source domain. This is referred to as initial marking or "pre-marking".

Consider the example of a company that has the policy that its CEO’s packets should have higher priority. The CEO’s host may mark the DS field of all outgoing packets with a DS codepoint that indicates "higher priority". Alternatively, the first-hop router directly connected to the CEO’s host may classify the traffic and mark the CEO’s packets with the correct DS codepoint. Such high priority traffic may also be conditioned near the source so that there is a limit on the amount of high priority traffic forwarded from a particular source.

There are some advantages to marking packets close to the traffic source. First, a traffic source can more easily take an application’s preferences into account when deciding which packets should receive better forwarding treatment. Also, classification of

packets is much simpler before the traffic has been aggregated with packets from other sources, since the number of classification rules which need to be applied within a single node is reduced.

Since packet marking may be distributed across multiple nodes, the source DS domain is responsible for ensuring that the aggregated traffic towards its provider DS domain conforms to the appropriate TCA. Additional allocation mechanisms such as bandwidth brokers or RSVP may be used to dynamically allocate resources for a particular DS behavior aggregate within the provider's network [2BIT, Bernet]. The boundary node of the source domain should also monitor conformance to the TCA, and may police, shape, or re-mark packets as necessary.

2.3.4.2 At the Boundary of a DS Domain

Traffic streams may be classified, marked, and otherwise conditioned on either end of a boundary link (the DS egress node of the upstream domain or the DS ingress node of the downstream domain). The SLA between the domains should specify which domain has responsibility for mapping traffic streams to DS behavior aggregates and conditioning those aggregates in conformance with the appropriate TCA. However, a DS ingress node must assume that the incoming traffic may not conform to the TCA and must be prepared to enforce the TCA in accordance with local policy.

When packets are pre-marked and conditioned in the upstream domain, potentially fewer classification and traffic conditioning rules need to be supported in the downstream DS domain. In this circumstance the downstream DS domain may only need to re-mark or police the incoming behavior aggregates to enforce the TCA. However, more sophisticated services which are path- or source-dependent may require MF classification in the downstream DS domain's ingress nodes.

If a DS ingress node is connected to an upstream non-DS-capable domain, the DS ingress node must be able to perform all necessary traffic conditioning functions on the incoming traffic.

2.3.4.3 In non-DS-Capable Domains

Traffic sources or intermediate nodes in a non-DS-capable domain may employ traffic conditioners to pre-mark traffic before it reaches the ingress of a downstream DS domain. In this way the local policies for classification and marking may be concealed.

2.3.4.4 In Interior DS Nodes

Although the basic architecture assumes that complex classification and traffic conditioning functions are located only in a network's ingress and egress boundary nodes, deployment of these functions in the interior of the network is not precluded. For example, more restrictive access policies may be enforced on a transoceanic link, requiring MF classification and traffic conditioning functionality in the upstream node on the link. This approach may have scaling limits, due to the potentially large number of classification and conditioning rules that might need to be maintained.

2.4 Per-Hop Behaviors

A per-hop behavior (PHB) is a description of the externally observable forwarding behavior of a DS node applied to a particular DS behavior aggregate. "Forwarding behavior" is a general concept in this context. For example, in the event that only one behavior aggregate occupies a link, the observable forwarding behavior (i.e., loss, delay, jitter) will often depend only on the relative loading of the link (i.e., in the event that the behavior assumes a work-conserving scheduling discipline). Useful behavioral distinctions are mainly observed when multiple behavior aggregates compete for buffer and bandwidth resources on a node. The PHB is the means by which a node allocates resources to behavior aggregates, and it is on top of this basic hop-by-hop resource allocation mechanism that useful differentiated services may be constructed.

The most simple example of a PHB is one which guarantees a minimal bandwidth allocation of X% of a link (over some reasonable time interval) to a behavior aggregate. This PHB can be fairly easily measured under a variety of competing traffic conditions. A slightly more complex PHB would guarantee a minimal bandwidth allocation of X% of a link, with proportional fair sharing of any excess link capacity. In general, the observable behavior of a PHB may depend on certain constraints on the traffic characteristics of the associated behavior aggregate, or the characteristics of other behavior aggregates.

PHBs may be specified in terms of their resource (e.g., buffer, bandwidth) priority relative to other PHBs, or in terms of their relative observable traffic characteristics (e.g., delay, loss). These PHBs may be used as building blocks to allocate resources and should be specified as a group (PHB group) for consistency. PHB groups will usually share a common constraint applying to each PHB within the group, such as a packet scheduling or buffer management policy. The relationship between PHBs in a group may be in terms of absolute or relative priority (e.g., discard priority by means of

deterministic or stochastic thresholds), but this is not required (e.g., N equal link shares). A single PHB defined in isolation is a special case of a PHB group.

PHBs are implemented in nodes by means of some buffer management and packet scheduling mechanisms. PHBs are defined in terms of behavior characteristics relevant to service provisioning policies, and not in terms of particular implementation mechanisms. In general, a variety of implementation mechanisms may be suitable for implementing a particular PHB group. Furthermore, it is likely that more than one PHB group may be implemented on a node and utilized within a domain. PHB groups should be defined such that the proper resource allocation between groups can be inferred, and integrated mechanisms can be implemented which can simultaneously support two or more groups. A PHB group definition should indicate possible conflicts with previously documented PHB groups which might prevent simultaneous operation.

As described in [DSFIELD], a PHB is selected at a node by a mapping of the DS codepoint in a received packet. Standardized PHBs have a recommended codepoint. However, the total space of codepoints is larger than the space available for recommended codepoints for standardized PHBs, and [DSFIELD] leaves provisions for locally configurable mappings. A codepoint->PHB mapping table may contain both 1->1 and N->1 mappings. All codepoints must be mapped to some PHB; in the absence of some local policy, codepoints which are not mapped to a standardized PHB in accordance with that PHB's specification should be mapped to the Default PHB.

2.5 Network Resource Allocation

The implementation, configuration, operation and administration of the supported PHB groups in the nodes of a DS Domain should effectively partition the resources of those nodes and the inter-node links between behavior aggregates, in accordance with the domain's service provisioning policy. Traffic conditioners can further control the usage of these resources through enforcement of TCAs and possibly through operational feedback from the nodes and traffic conditioners in the domain. Although a range of services can be deployed in the absence of complex traffic conditioning functions (e.g., using only static marking policies), functions such as policing, shaping, and dynamic re-marking enable the deployment of services providing quantitative performance metrics.

The configuration of and interaction between traffic conditioners and interior nodes should be managed by the administrative control of the domain and may require operational control through protocols and a control entity. There is a wide range of possible control models.

The precise nature and implementation of the interaction between these components is outside the scope of this architecture. However, scalability requires that the control of the domain does not require micro-management of the network resources. The most scalable control model would operate nodes in open-loop in the operational timeframe, and would only require administrative-timescale management as SLAs are varied. This simple model may be unsuitable in some circumstances, and some automated but slowly varying operational control (minutes rather than seconds) may be desirable to balance the utilization of the network against the recent load profile.

3. Per-Hop Behavior Specification Guidelines

Basic requirements for per-hop behavior standardization are given in [DSFIELD]. This section elaborates on that text by describing additional guidelines for PHB (group) specifications. This is intended to help foster implementation consistency. Before a PHB group is proposed for standardization it should satisfy these guidelines, as appropriate, to preserve the integrity of this architecture.

G.1: A PHB standard must specify a recommended DS codepoint selected from the codepoint space reserved for standard mappings [DSFIELD]. Recommended codepoints will be assigned by the IANA. A PHB proposal may recommend a temporary codepoint from the EXP/LU space to facilitate inter-domain experimentation. Determination of a packet's PHB must not require inspection of additional packet header fields beyond the DS field.

G.2: The specification of each newly proposed PHB group should include an overview of the behavior and the purpose of the behavior being proposed. The overview should include a problem or problems statement for which the PHB group is targeted. The overview should include the basic concepts behind the PHB group. These concepts should include, but are not restricted to, queueing behavior, discard behavior, and output link selection behavior. Lastly, the overview should specify the method by which the PHB group solves the problem or problems specified in the problem statement.

G.3: A PHB group specification should indicate the number of individual PHBs specified. In the event that multiple PHBs are specified, the interactions between these PHBs and constraints that must be respected globally by all the PHBs within the group should be clearly specified. As an example, the specification must indicate whether the probability of packet reordering within a microflow is increased if different packets in that microflow are marked for different PHBs within the group.

G.4: When proper functioning of a PHB group is dependent on constraints such as a provisioning restriction, then the PHB definition should describe the behavior when these constraints are violated. Further, if actions such as packet discard or re-marking are required when these constraints are violated, then these actions should be specifically stipulated.

G.5: A PHB group may be specified for local use within a domain in order to provide some domain-specific functionality or domain-specific services. In this event, the PHB specification is useful for providing vendors with a consistent definition of the PHB group. However, any PHB group which is defined for local use should not be considered for standardization, but may be published as an Informational RFC. In contrast, a PHB group which is intended for general use will follow a stricter standardization process. Therefore all PHB proposals should specifically state whether they are to be considered for general or local use.

It is recognized that PHB groups can be designed with the intent of providing host-to-host, WAN edge-to-WAN edge, and/or domain edge-to-domain edge services. Use of the term "end-to-end" in a PHB definition should be interpreted to mean "host-to-host" for consistency.

Other PHB groups may be defined and deployed locally within domains, for experimental or operational purposes. There is no requirement that these PHB groups must be publicly documented, but they should utilize DS codepoints from one of the EXP/LU pools as defined in [DSFIELD].

G.6: It may be possible or appropriate for a packet marked for a PHB within a PHB group to be re-marked to select another PHB within the group; either within a domain or across a domain boundary. Typically there are three reasons for such PHB modification:

a. The codepoints associated with the PHB group are collectively intended to carry state about the network,
b. Conditions exist which require PHB promotion or demotion of a packet (this assumes that PHBs within the group can be ranked in some order),
c. The boundary between two domains is not covered by a SLA. In this case the codepoint/PHB to select when crossing the boundary link will be determined by the local policy of the upstream domain.

A PHB specification should clearly state the circumstances under which packets marked for a PHB within a PHB group may, or should be modified (e.g., promoted or demoted) to another PHB within the group. If it is undesirable for a packet’s PHB to be modified, the

specification should clearly state the consequent risks when the PHB is modified. A possible risk to changing a packet's PHB, either within or outside a PHB group, is a higher probability of packet re-ordering within a microflow. PHBs within a group may carry some host-to-host, WAN edge-to-WAN edge, and/or domain edge-to-domain edge semantics which may be difficult to duplicate if packets are re-marked to select another PHB from the group (or otherwise).

For certain PHB groups, it may be appropriate to reflect a state change in the node by re-marking packets to specify another PHB from within the group. If a PHB group is designed to reflect the state of a network, the PHB definition must adequately describe the relationship between the PHBs and the states they reflect. Further, if these PHBs limit the forwarding actions a node can perform in some way, these constraints may be specified as actions the node should, or must perform.

G.7: A PHB group specification should include a section defining the implications of tunneling on the utility of the PHB group. This section should specify the implications for the utility of the PHB group of a newly created outer header when the original DS field of the inner header is encapsulated in a tunnel. This section should also discuss what possible changes should be applied to the inner header at the egress of the tunnel, when both the codepoints from the inner header and the outer header are accessible (see Sec. 6.2).

G.8: The process of specifying PHB groups is likely to be incremental in nature. When new PHB groups are proposed, their known interactions with previously specified PHB groups should be documented. When a new PHB group is created, it can be entirely new in scope or it can be an extension to an existing PHB group. If the PHB group is entirely independent of some or all of the existing PHB specifications, a section should be included in the PHB specification which details how the new PHB group can co-exist with those PHB groups already standardized. For example, this section might indicate the possibility of packet re-ordering within a microflow for packets marked by codepoints associated with two separate PHB groups. If concurrent operation of two (or more) different PHB groups in the same node is impossible or detrimental this should be stated. If the concurrent operation of two (or more) different PHB groups requires some specific behaviors by the node when packets marked for PHBs from these different PHB groups are being processed by the node at the same time, these behaviors should be stated.

Care should be taken to avoid circularity in the definitions of PHB groups.

If the proposed PHB group is an extension to an existing PHB group, a section should be included in the PHB group specification which details how this extension interoperates with the behavior being extended. Further, if the extension alters or more narrowly defines the existing behavior in some way, this should also be clearly indicated.

G.9: Each PHB specification should include a section specifying minimal conformance requirements for implementations of the PHB group. This conformance section is intended to provide a means for specifying the details of a behavior while allowing for implementation variation to the extent permitted by the PHB specification. This conformance section can take the form of rules, tables, pseudo-code, or tests.

G.10: A PHB specification should include a section detailing the security implications of the behavior. This section should include a discussion of the re-marking of the inner header's codepoint at the egress of a tunnel and its effect on the desired forwarding behavior.

Further, this section should also discuss how the proposed PHB group could be used in denial-of-service attacks, reduction of service contract attacks, and service contract violation attacks. Lastly, this section should discuss possible means for detecting such attacks as they are relevant to the proposed behavior.

G.11: A PHB specification should include a section detailing configuration and management issues which may affect the operation of the PHB and which may impact candidate services that might utilize the PHB.

G.12: It is strongly recommended that an appendix be provided with each PHB specification that considers the implications of the proposed behavior on current and potential services. These services could include but are not restricted to be user-specific, device-specific, domain-specific or end-to-end services. It is also strongly recommended that the appendix include a section describing how the services are verified by users, devices, and/or domains.

G.13: It is recommended that an appendix be provided with each PHB specification that is targeted for local use within a domain, providing guidance for PHB selection for packets which are forwarded into a peer domain which does not support the PHB group.

G.14: It is recommended that an appendix be provided with each PHB specification which considers the impact of the proposed PHB group on existing higher-layer protocols. Under some circumstances PHBs may allow for possible changes to higher-layer protocols which may increase or decrease the utility of the proposed PHB group.

G.15: It is recommended that an appendix be provided with each PHB specification which recommends mappings to link-layer QoS mechanisms to support the intended behavior of the PHB across a shared-medium or switched link-layer. The determination of the most appropriate mapping between a PHB and a link-layer QoS mechanism is dependent on many factors and is outside the scope of this document; however, the specification should attempt to offer some guidance.

4. Interoperability with Non-Differentiated Services-Compliant Nodes

We define a non-differentiated services-compliant node (non-DS-compliant node) as any node which does not interpret the DS field as specified in [DSFIELD] and/or does not implement some or all of the standardized PHBs (or those in use within a particular DS domain). This may be due to the capabilities or configuration of the node. We define a legacy node as a special case of a non-DS-compliant node which implements IPv4 Precedence classification and forwarding as defined in [RFC791, RFC1812], but which is otherwise not DS-compliant. The precedence values in the IPv4 TOS octet are compatible by intention with the Class Selector Codepoints defined in [DSFIELD], and the precedence forwarding behaviors defined in [RFC791, RFC1812] comply with the Class Selector PHB Requirements also defined in [DSFIELD]. A key distinction between a legacy node and a DS-compliant node is that the legacy node may or may not interpret bits 3-6 of the TOS octet as defined in [RFC1349] (the "DTRC" bits); in practice it will not interpret these bit as specified in [DSFIELD]. We assume that the use of the TOS markings defined in [RFC1349] is deprecated. Nodes which are non-DS-compliant and which are not legacy nodes may exhibit unpredictable forwarding behaviors for packets with non-zero DS codepoints.

Differentiated services depend on the resource allocation mechanisms provided by per-hop behavior implementations in nodes. The quality or statistical assurance level of a service may break down in the event that traffic transits a non-DS-compliant node, or a non-DS-capable domain.

We will examine two separate cases. The first case concerns the use of non-DS-compliant nodes within a DS domain. Note that PHB forwarding is primarily useful for allocating scarce node and link resources in a controlled manner. On high-speed, lightly loaded links, the worst-case packet delay, jitter, and loss may be

negligible, and the use of a non-DS-compliant node on the upstream end of such a link may not result in service degradation. In more realistic circumstances, the lack of PHB forwarding in a node may make it impossible to offer low-delay, low-loss, or provisioned bandwidth services across paths which traverse the node. However, use of a legacy node may be an acceptable alternative, assuming that the DS domain restricts itself to using only the Class Selector Codepoints defined in [DSFIELD], and assuming that the particular precedence implementation in the legacy node provides forwarding behaviors which are compatible with the services offered along paths which traverse that node. Note that it is important to restrict the codepoints in use to the Class Selector Codepoints, since the legacy node may or may not interpret bits 3-5 in accordance with [RFC1349], thereby resulting in unpredictable forwarding results.

The second case concerns the behavior of services which traverse non-DS-capable domains. We assume for the sake of argument that a non-DS-capable domain does not deploy traffic conditioning functions on domain boundary nodes; therefore, even in the event that the domain consists of legacy or DS-compliant interior nodes, the lack of traffic enforcement at the boundaries will limit the ability to consistently deliver some types of services across the domain. A DS domain and a non-DS-capable domain may negotiate an agreement which governs how egress traffic from the DS-domain should be marked before entry into the non-DS-capable domain. This agreement might be monitored for compliance by traffic sampling instead of by rigorous traffic conditioning. Alternatively, where there is knowledge that the non-DS-capable domain consists of legacy nodes, the upstream DS domain may opportunistically re-mark differentiated services traffic to one or more of the Class Selector Codepoints. Where there is no knowledge of the traffic management capabilities of the downstream domain, and no agreement in place, a DS domain egress node may choose to re-mark DS codepoints to zero, under the assumption that the non-DS-capable domain will treat the traffic uniformly with best-effort service.

In the event that a non-DS-capable domain peers with a DS domain, traffic flowing from the non-DS-capable domain should be conditioned at the DS ingress node of the DS domain according to the appropriate SLA or policy.

5. Multicast Considerations

Use of differentiated services by multicast traffic introduces a number of issues for service provisioning. First, multicast packets which enter a DS domain at an ingress node may simultaneously take multiple paths through some segments of the domain due to multicast packet replication. In this way they consume more network resources

than unicast packets. Where multicast group membership is dynamic, it is difficult to predict in advance the amount of network resources that may be consumed by multicast traffic originating from an upstream network for a particular group. A consequence of this uncertainty is that it may be difficult to provide quantitative service guarantees to multicast senders. Further, it may be necessary to reserve codepoints and PHBs for exclusive use by unicast traffic, to provide resource isolation from multicast traffic.

The second issue is the selection of the DS codepoint for a multicast packet arriving at a DS ingress node. Because that packet may exit the DS domain at multiple DS egress nodes which peer with multiple downstream domains, the DS codepoint used should not result in the request for a service from a downstream DS domain which is in violation of a peering SLA. When establishing classifier and traffic conditioner state at an DS ingress node for an aggregate of traffic receiving a differentiated service which spans across the egress boundary of the domain, the identity of the adjacent downstream transit domain and the specifics of the corresponding peering SLA can be factored into the configuration decision (subject to routing policy and the stability of the routing infrastructure). In this way peering SLAs with downstream DS domains can be partially enforced at the ingress of the upstream domain, reducing the classification and traffic conditioning burden at the egress node of the upstream domain. This is not so easily performed in the case of multicast traffic, due to the possibility of dynamic group membership. The result is that the service guarantees for unicast traffic may be impacted. One means of addressing this problem is to establish a separate peering SLA for multicast traffic, and to either utilize a particular set of codepoints for multicast packets, or to implement the necessary classification and traffic conditioning mechanisms in the DS egress nodes to provide preferential isolation for unicast traffic in conformance with the peering SLA with the downstream domain.

6. Security and Tunneling Considerations

This section addresses security issues raised by the introduction of differentiated services, primarily the potential for denial-of-service attacks, and the related potential for theft of service by unauthorized traffic (Sec. 6.1). In addition, the operation of differentiated services in the presence of IPsec and its interaction with IPsec are also discussed (Sec. 6.2), as well as auditing requirements (Sec. 6.3). This section considers issues introduced by the use of both IPsec and non-IPsec tunnels.

6.1 Theft and Denial of Service

The primary goal of differentiated services is to allow different levels of service to be provided for traffic streams on a common network infrastructure. A variety of resource management techniques may be used to achieve this, but the end result will be that some packets receive different (e.g., better) service than others. The mapping of network traffic to the specific behaviors that result in different (e.g., better or worse) service is indicated primarily by the DS field, and hence an adversary may be able to obtain better service by modifying the DS field to codepoints indicating behaviors used for enhanced services or by injecting packets with the DS field set to such codepoints. Taken to its limits, this theft of service becomes a denial-of-service attack when the modified or injected traffic depletes the resources available to forward it and other traffic streams. The defense against such theft- and denial-of-service attacks consists of the combination of traffic conditioning at DS boundary nodes along with security and integrity of the network infrastructure within a DS domain.

As described in Sec. 2, DS ingress nodes must condition all traffic entering a DS domain to ensure that it has acceptable DS codepoints. This means that the codepoints must conform to the applicable TCA(s) and the domain's service provisioning policy. Hence, the ingress nodes are the primary line of defense against theft- and denial-of-service attacks based on modified DS codepoints (e.g., codepoints to which the traffic is not entitled), as success of any such attack constitutes a violation of the applicable TCA(s) and/or service provisioning policy. An important instance of an ingress node is that any traffic-originating node in a DS domain is the ingress node for that traffic, and must ensure that all originated traffic carries acceptable DS codepoints.

Both a domain's service provisioning policy and TCAs may require the ingress nodes to change the DS codepoint on some entering packets (e.g., an ingress router may set the DS codepoint of a customer's traffic in accordance with the appropriate SLA). Ingress nodes must condition all other inbound traffic to ensure that the DS codepoints are acceptable; packets found to have unacceptable codepoints must either be discarded or must have their DS codepoints modified to acceptable values before being forwarded. For example, an ingress node receiving traffic from a domain with which no enhanced service agreement exists may reset the DS codepoint to the Default PHB codepoint [DSFIELD]. Traffic authentication may be required to validate the use of some DS codepoints (e.g., those corresponding to enhanced services), and such authentication may be performed by technical means (e.g., IPsec) and/or non-technical means (e.g., the inbound link is known to be connected to exactly one customer site).

An inter-domain agreement may reduce or eliminate the need for ingress node traffic conditioning by making the upstream domain partly or completely responsible for ensuring that traffic has DS codepoints acceptable to the downstream domain. In this case, the ingress node may still perform redundant traffic conditioning checks to reduce the dependence on the upstream domain (e.g., such checks can prevent theft-of-service attacks from propagating across the domain boundary). If such a check fails because the upstream domain is not fulfilling its responsibilities, that failure is an auditable event; the generated audit log entry should include the date/time the packet was received, the source and destination IP addresses, and the DS codepoint that caused the failure. In practice, the limited gains from such checks need to be weighed against their potential performance impact in determining what, if any, checks to perform under these circumstances.

Interior nodes in a DS domain may rely on the DS field to associate differentiated services traffic with the behaviors used to implement enhanced services. Any node doing so depends on the correct operation of the DS domain to prevent the arrival of traffic with unacceptable DS codepoints. Robustness concerns dictate that the arrival of packets with unacceptable DS codepoints must not cause the failure (e.g., crash) of network nodes. Interior nodes are not responsible for enforcing the service provisioning policy (or individual SLAs) and hence are not required to check DS codepoints before using them. Interior nodes may perform some traffic conditioning checks on DS codepoints (e.g., check for DS codepoints that are never used for traffic on a specific link) to improve security and robustness (e.g., resistance to theft-of-service attacks based on DS codepoint modifications). Any detected failure of such a check is an auditable event and the generated audit log entry should include the date/time the packet was received, the source and destination IP addresses, and the DS codepoint that caused the failure. In practice, the limited gains from such checks need to be weighed against their potential performance impact in determining what, if any, checks to perform at interior nodes.

Any link that cannot be adequately secured against modification of DS codepoints or traffic injection by adversaries should be treated as a boundary link (and hence any arriving traffic on that link is treated as if it were entering the domain at an ingress node). Local security policy provides the definition of "adequately secured," and such a definition may include a determination that the risks and consequences of DS codepoint modification and/or traffic injection do not justify any additional security measures for a link. Link security can be enhanced via physical access controls and/or software means such as tunnels that ensure packet integrity.

6.2 IPsec and Tunneling Interactions

The IPsec protocol, as defined in [ESP, AH], does not include the IP header's DS field in any of its cryptographic calculations (in the case of tunnel mode, it is the outer IP header's DS field that is not included). Hence modification of the DS field by a network node has no effect on IPsec's end-to-end security, because it cannot cause any IPsec integrity check to fail. As a consequence, IPsec does not provide any defense against an adversary's modification of the DS field (i.e., a man-in-the-middle attack), as the adversary's modification will also have no effect on IPsec's end-to-end security. In some environments, the ability to modify the DS field without affecting IPsec integrity checks may constitute a covert channel; if it is necessary to eliminate such a channel or reduce its bandwidth, the DS domains should be configured so that the required processing (e.g., set all DS fields on sensitive traffic to a single value) can be performed at DS egress nodes where traffic exits higher security domains.

IPsec's tunnel mode provides security for the encapsulated IP header's DS field. A tunnel mode IPsec packet contains two IP headers: an outer header supplied by the tunnel ingress node and an encapsulated inner header supplied by the original source of the packet. When an IPsec tunnel is hosted (in whole or in part) on a differentiated services network, the intermediate network nodes operate on the DS field in the outer header. At the tunnel egress node, IPsec processing includes stripping the outer header and forwarding the packet (if required) using the inner header. If the inner IP header has not been processed by a DS ingress node for the tunnel egress node's DS domain, the tunnel egress node is the DS ingress node for traffic exiting the tunnel, and hence must carry out the corresponding traffic conditioning responsibilities (see Sec. 6.1). If the IPsec processing includes a sufficiently strong cryptographic integrity check of the encapsulated packet (where sufficiency is determined by local security policy), the tunnel egress node can safely assume that the DS field in the inner header has the same value as it had at the tunnel ingress node. This allows a tunnel egress node in the same DS domain as the tunnel ingress node, to safely treat a packet passing such an integrity check as if it had arrived from another node within the same DS domain, omitting the DS ingress node traffic conditioning that would otherwise be required. An important consequence is that otherwise insecure links internal to a DS domain can be secured by a sufficiently strong IPsec tunnel.

This analysis and its implications apply to any tunneling protocol that performs integrity checks, but the level of assurance of the inner header's DS field depends on the strength of the integrity

check performed by the tunneling protocol. In the absence of sufficient assurance for a tunnel that may transit nodes outside the current DS domain (or is otherwise vulnerable), the encapsulated packet must be treated as if it had arrived at a DS ingress node from outside the domain.

The IPsec protocol currently requires that the inner header's DS field not be changed by IPsec decapsulation processing at a tunnel egress node. This ensures that an adversary's modifications to the DS field cannot be used to launch theft- or denial-of-service attacks across an IPsec tunnel endpoint, as any such modifications will be discarded at the tunnel endpoint. This document makes no change to that IPsec requirement.

If the IPsec specifications are modified in the future to permit a tunnel egress node to modify the DS field in an inner IP header based on the DS field value in the outer header (e.g., copying part or all of the outer DS field to the inner DS field), then additional considerations would apply. For a tunnel contained entirely within a single DS domain and for which the links are adequately secured against modifications of the outer DS field, the only limits on inner DS field modifications would be those imposed by the domain's service provisioning policy. Otherwise, the tunnel egress node performing such modifications would be acting as a DS ingress node for traffic exiting the tunnel and must carry out the traffic conditioning responsibilities of an ingress node, including defense against theft- and denial-of-service attacks (See Sec. 6.1). If the tunnel enters the DS domain at a node different from the tunnel egress node, the tunnel egress node may depend on the upstream DS ingress node having ensured that the outer DS field values are acceptable. Even in this case, there are some checks that can only be performed by the tunnel egress node (e.g., a consistency check between the inner and outer DS codepoints for an encrypted tunnel). Any detected failure of such a check is an auditable event and the generated audit log entry should include the date/time the packet was received, the source and destination IP addresses, and the DS codepoint that was unacceptable.

An IPsec tunnel can be viewed in at least two different ways from an architectural perspective. If the tunnel is viewed as a logical single hop "virtual wire", the actions of intermediate nodes in forwarding the tunneled traffic should not be visible beyond the ends of the tunnel and hence the DS field should not be modified as part of decapsulation processing. In contrast, if the tunnel is viewed as a multi-hop participant in forwarding traffic, then modification of the DS field as part of tunnel decapsulation processing may be desirable. A specific example of the latter situation occurs when a tunnel terminates at an interior node of a DS domain at which the domain administrator does not wish to deploy traffic conditioning

logic (e.g., to simplify traffic management). This could be supported by using the DS codepoint in the outer IP header (which was subject to traffic conditioning at the DS ingress node) to reset the DS codepoint in the inner IP header, effectively moving DS ingress traffic conditioning responsibilities from the IPsec tunnel egress node to the appropriate upstream DS ingress node (which must already perform that function for unencapsulated traffic).

6.3 Auditing

Not all systems that support differentiated services will implement auditing. However, if differentiated services support is incorporated into a system that supports auditing, then the differentiated services implementation should also support auditing. If such support is present the implementation must allow a system administrator to enable or disable auditing for differentiated services as a whole, and may allow such auditing to be enabled or disabled in part.

For the most part, the granularity of auditing is a local matter. However, several auditable events are identified in this document and for each of these events a minimum set of information that should be included in an audit log is defined. Additional information (e.g., packets related to the one that triggered the auditable event) may also be included in the audit log for each of these events, and additional events, not explicitly called out in this specification, also may result in audit log entries. There is no requirement for the receiver to transmit any message to the purported sender in response to the detection of an auditable event, because of the potential to induce denial of service via such action.

7. Acknowledgements

This document has benefitted from earlier drafts by Steven Blake, David Clark, Ed Ellesson, Paul Ferguson, Juha Heinanen, Van Jacobson, Kalevi Kilkki, Kathleen Nichols, Walter Weiss, John Wroclawski, and Lixia Zhang.

The authors would like to acknowledge the following individuals for their helpful comments and suggestions: Kathleen Nichols, Brian Carpenter, Konstantinos Dovrolis, Shivkumar Kalyana, Wu-chang Feng, Marty Borden, Yoram Bernet, Ronald Bonica, James Binder, Borje Ohlman, Alessio Casati, Scott Brim, Curtis Villamizar, Hamid Ould-Brahi, Andrew Smith, John Renwick, Werner Almesberger, Alan O'Neill, James Fu, and Bob Braden.

8. References

[802.1p] ISO/IEC Final CD 15802-3 Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Common specifications - Part 3: Media Access Control (MAC) bridges, (current draft available as IEEE P802.1D/D15).

[AH] Kent, S. and R. Atkinson, "IP Authentication Header", RFC 2402, November 1998.

[ATM] ATM Traffic Management Specification Version 4.0 <af-tm-0056.000>, ATM Forum, April 1996.

[Bernet] Y. Bernet, R. Yavatkar, P. Ford, F. Baker, L. Zhang, K. Nichols, and M. Speer, "A Framework for Use of RSVP with Diff-serv Networks", Work in Progress.

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

[EXPLICIT] D. Clark and W. Fang, "Explicit Allocation of Best Effort Packet Delivery Service", IEEE/ACM Trans. on Networking, vol. 6, no. 4, August 1998, pp. 362-373.

[ESP] Kent, S. and R. Atkinson, "IP Encapsulating Security Payload (ESP)", RFC 2406, November 1998.

[FRELAY] ANSI T1S1, "DSSI Core Aspects of Frame Rely", March 1990.

[RFC791] Postel, J., Editor, "Internet Protocol", STD 5, RFC 791, September 1981.

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

[RFC1633] Braden, R., Clark, D. and S. Shenker, "Integrated Services in the Internet Architecture: An Overview", RFC 1633, July 1994.

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

[RSVP] Braden, B., Zhang, L., Berson S., Herzog, S. and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.

[2BIT] K. Nichols, V. Jacobson, and L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet", ftp://ftp.ee.lbl.gov/papers/dsarch.pdf, November 1997.

[TR] ISO/IEC 8802-5 Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Common specifications - Part 5: Token Ring Access Method and Physical Layer Specifications, (also ANSI/IEEE Std 802.5-1995), 1995.

Authors' Addresses

Steven Blake
Torrent Networking Technologies
3000 Aerial Center, Suite 140
Morrisville, NC 27560

Phone: +1-919-468-8466 x232
EMail: slblake@torrentnet.com

David L. Black
EMC Corporation
35 Parkwood Drive
Hopkinton, MA 01748

Phone: +1-508-435-1000 x76140
EMail: black_david@emc.com

Mark A. Carlson
Sun Microsystems, Inc.
2990 Center Green Court South
Boulder, CO 80301

Phone: +1-303-448-0048 x115
EMail: mark.carlson@sun.com

Elwyn Davies
Nortel UK
London Road
Harlow, Essex CM17 9NA, UK

Phone: +44-1279-405498
EMail: elwynd@nortel.co.uk

Zheng Wang
Bell Labs Lucent Technologies
101 Crawfords Corner Road
Holmdel, NJ 07733

EMail: zhwang@bell-labs.com

Walter Weiss
Lucent Technologies
300 Baker Avenue, Suite 100
Concord, MA 01742-2168

EMail: wweiss@lucent.com

Full Copyright Statement