IP Quality-of-Service

Kimmo Raatikainen

Presentation Outline

- IP QoS Architectures
- Next Steps in IP QoS
- Resource Reservation Protocol (RSVP)
- Differentiated Services

IP QoS Architectures

- Integrated Services (RFC1633)
  - state-based
  - explicit reservation (RSVP)
  - quaranted service
- Differentiated Services (RFC2475)
  - stateless
  - no reservation
  - better than best effort service

Next Steps for QoS Architecture

- RFC2990
- Aggregation of state in IntServ
- IntServ over DiffServ networks (RFC2998)
- Per-Domain Behavior in DiffServ
- Service Level Agreements between DiffServ domains
- Multiprotocol Label Switching

Resource ReSerVation Protocol (RSVP)

Version 1 Functional Specification

RFC 2205

September 1997

RSVP in Nutshell - 1/3

- RSVP makes resource reservations for both unicast and many-to-many multicast applications, adapting dynamically to changing group membership as well as to changing routes.
- RSVP is simplex, i.e., it makes reservations for unidirectional data flows.
- RSVP is receiver-oriented, i.e., the receiver of a data flow initiates and maintains the resource reservation used for that flow.

RSVP in Nutshell - 2/3

- RSVP maintains "soft" state in routers and hosts, providing graceful support for dynamic membership changes and automatic adaptation to routing changes.
- RSVP is not a routing protocol but depends upon present and future routing protocols.
- RSVP transports and maintains traffic control and policy control parameters that are opaque to RSVP.

RSVP in Nutshell - 3/3

- RSVP provides several reservation models or "styles" (defined below) to fit a variety of applications.
- RSVP provides transparent operation through routers that do not support it.
- RSVP supports both IPv4 and IPv6.

RSVP: Introduction - 1/6

- RSVP is a resource reservation setup protocol designed for an integrated services Internet [RSVP93, RFC 1633].
- The RSVP protocol is used
  - by a host to request specific qualities of service from the network for particular application data streams or flows.
  - by routers to deliver quality-of-service (QoS) requests to all nodes along the path(s) of the flows and to establish and maintain state to provide the requested service.

RSVP: Introduction - 2/6

- RSVP requests will generally result in resources being reserved in each node along the data path.
- RSVP requests resources for simplex flows, i.e., it requests resources in only one direction.
- Therefore, RSVP treats a sender as logically distinct from a receiver, although the same application process may act as both a sender and a receiver at the same time.

RSVP: Introduction - 3/6

- RSVP operates on top of IPv4 or IPv6, occupying the place of a transport protocol in the protocol stack.
- RSVP does not transport application data!
- It is rather an Internet control protocol, like ICMP, IGMP, or routing protocols.
- Like the implementations of routing and management protocols, an implementation of RSVP will typically execute in the background.

RSVP in Hosts and Routers

```
                HOST                              ROUTER

  _____________________________       ____________________________
 |  _______                    |     |                            |
 | |       |   _______         |     |            _______         |
 | |Appli- |  |       |        |RSVP |           |       |        |
 | | cation|  | RSVP <---------------------------> RSVP  <----------->
 | |       <-->       |        |     | _______   |       |        |
 | |       |  |process|  _____ |     ||Routing|  |process|  _____ |
 | |_._____|  |       -->Polcy||     ||       <-->       -->Polcy||
 |   |        |__.__._| |Cntrl||     ||process|  |__.__._| |Cntrl||
 |   |data       |  |   |_____||     ||__.____|     |  |   |_____||
 |===|===========|==|==========|     |===|==========|==|==========|
 |   |   --------|  |    _____ |     |   |  --------|  |    _____ |
 |   |  |        |  ---->Admis||     |   |  |       |  ---->Admis||
 |  _V__V_    ___V____  |Cntrl||     |  _V__V_    __V_____ |Cntrl||
 | |      |  |        | |_____||     | |      |  |        ||_____||
 | |Class-|  | Packet |        |     | |Class-|  | Packet |       |
 | | ifier|==>Schedulr|================> ifier|==>Schedulr|===========>
 | |______|  |________|        |data | |______|  |________|       |data
 |                             |     |                            |
 |_____________________________|     |____________________________|
```

RSVP: Introduction - 4/6

- RSVP is not itself a routing protocol
  - designed to operate with current and future unicast and multicast routing protocols
- An RSVP process consults the local routing database(s) to obtain routes.
- Routing protocols determine where packets get forwarded
- RSVP is only concerned with the QoS of those packets that are forwarded in accordance with routing.

RSVP: Introduction - 5/6

- In order to efficiently accommodate
  - large groups,
  - dynamic group membership,
  - and heterogeneous receiver requirements,

RSVP makes receivers responsible for requesting a specific QoS.

RSVP: Introduction - 6/6

- A QoS request from a receiver host application is passed to the local RSVP process.
- The RSVP protocol then carries the request to all the nodes (routers and hosts) along the reverse data path(s) to the data source(s), but only as far as the router where the receiver's data path joins the multicast distribution tree.
- As a result, RSVP's reservation overhead is in general logarithmic rather than linear in the number of receivers.

Traffic Control - 1/2

- Quality of service is implemented for a particular data flow by mechanisms collectively called "traffic control".
- These mechanisms include
  - a packet classifier,
  - admission control, and
  - a "packet scheduler”
    - or some other link-layer-dependent mechanism to determine when particular packets are forwarded.

Traffic Control - 2/2

- The "packet classifier" determines the QoS class (and perhaps the route) for each packet.
- For each outgoing interface, the "packet scheduler" achieves the promised QoS.
- Traffic control implements QoS service models defined by the Integrated Services Working Group.

Reservation Setup - 1/2

- During reservation setup, an RSVP QoS request is passed to two local decision modules, "admission control" and "policy control".
- Admission control determines whether the node has sufficient available resources to supply the requested QoS.
- Policy control determines whether the user has administrative permission to make the reservation.

Reservation Setup - 2/2

- If both checks succeed,
  - parameters are set in the packet classifier and in the link layer interface (e.g., in the packet scheduler) to obtain the desired QoS.
- If either check fails,
  - the RSVP program returns an error notification to the application process that originated the request.

Reservation State - 1/2

- RSVP protocol mechanisms provide a general facility for creating and maintaining distributed reservation state across a mesh of multicast or unicast delivery paths.
- RSVP itself transfers and manipulates QoS and policy control parameters as opaque data, passing them to the appropriate traffic control and policy control modules for interpretation.

Reservation State - 2/2

- Since the membership of a large multicast group and the resulting multicast tree topology are likely to change with time,
  - the RSVP design assumes that state for RSVP and traffic control state is to be built and destroyed incrementally in routers and hosts.
- RSVP establishes "soft" state:
  - RSVP sends periodic refresh messages to maintain the state along the reserved path(s).
  - In the absence of refresh messages, the state automatically times out and is deleted.

Data Flows - 1/2

- RSVP defines a "session" to be a data flow with a particular destination and transport-layer protocol.
- RSVP treats each session independently, and this document often omits the implied qualification "for the same session".
- An RSVP session is defined by the triple:
  - (DestAddress, ProtocolId [, DstPort]).

Data Flows - 2/2

- It is not strictly necessary to include DstPort in the session definition when DestAddress is multicast
  - different sessions can always have different multicast addresses.
- DstPort is necessary to allow more than one unicast session addressed to the same receiver host.

Reservation Styles - 1/2

- One reservation option concerns the treatment of reservations for different senders within the same session:
  - establish a "distinct" reservation for each upstream sender, or
  - else make a single reservation that is "shared" among all packets of selected senders.

Reservation Styles - 2/2

- Another reservation option controls the selection of senders;
  - it may be an "explicit" list of all selected senders, or
  - a "wildcard" that implicitly selects all the senders to the session.
- In an explicit sender-selection reservation, each filter spec must match exactly one sender, while in a wildcard sender-selection no filter spec is needed.

Reservation Attributes and Styles

Sender Selection	Reservations: Distinct	Shared
Explicit	Fixed-Filter (FF) style	Shared-Explicit (SE) Style
Wildcard	(None defined)	Wildcard-Filter (WF) Style

Wildcard-Filter (WF) Style - 1/2

- implies the options:
  - "shared" reservation and
  - "wildcard" sender selection.
- AWF-style reservation creates a single reservation shared by flows from all upstream senders.
- This reservation may be thought of as a shared "pipe", whose "size" is the largest of the resource requests from all receivers, independent of the number of senders using it.

Wildcard-Filter (WF) Style - 2/2

- A WF-style reservation is propagated upstream towards all sender hosts.
- It automatically extends to new senders as they appear.

Fixed-Filter (FF) Style - 1/2

- implies the options:
  - "distinct" reservations and
  - "explicit" sender selection.
- An elementary FF-style reservation request creates a distinct reservation for data packets from a particular sender, not sharing them with other senders' packets for the same session.

Fixed-Filter (FF) Style - 2/2

- RSVP allows multiple elementary FF-style reservations to be requested at the same time, using a list of flow descriptors.
- The total reservation on a link for a given session is the `sum' of Q1, Q2, ... for all requested senders.

Shared Explicit (SE) Style

- implies the options:
  - "shared" reservation and
  - "explicit" sender selection.
- An SE-style reservation creates a single reservation shared by selected upstream senders.
- Unlike the WF style, the SE style allows a receiver to explicitly specify the set of senders to be included.

RSVP Messages - 1/2

- There are two fundamental RSVP message types:
  - Resv and
  - Path.
- Each receiver host sends RSVP reservation request (Resv) messages upstream towards the senders.
- Each RSVP sender host transmits RSVP "Path" messages downstream along the uni-/multicast routes provided by the routing protocol(s), following the paths of the data.

Resv Messages

- These messages must follow exactly the reverse of the path(s) the data packets will use, upstream to all the sender hosts included in the sender selection.
- They create and maintain "reservation state" in each node along the path(s).
- Resv messages must finally be delivered to the sender hosts themselves, so that the hosts can set up appropriate traffic control parameters for the first hop.

Path Messages - 1/2

- These Path messages store "path state" in each node along the way.
- This path state includes at least the unicast IP address of the previous hop node, which is used to route the Resv messages hop-by-hop in the reverse direction.

Path Messages - 2/2

- A Path message contains the following information in addition to the previous hop address:
  - Sender Template
  - Sender Tspec
  - Adspec

Sender Template - 1/2

- A Path message is required to carry a Sender Template, which describes the format of data packets that the sender will originate.
- This template is in the form of a filter spec that could be used to select this sender's packets from others in the same session on the same link.
- Sender Templates have exactly the same expressive power and format as filter specs that appear in Resv messages.

Sender Template - 2/2

- A a Sender Template may specify only the sender IP address and optionally the UDP/TCP sender port.
- It assumes the protocol Id specified for the session.

Sender Tspec

- A Path message is required to carry a Sender Tspec, which defines the traffic characteristics of the data flow that the sender will generate.
- This Tspec is used by traffic control to prevent over-reservation, and perhaps unnecessary Admission Control failures.

Adspec

- A Path message may carry a package of OPWA advertising information, known as an "Adspec".
- An Adspec received in a Path message
  - is passed to the local traffic control,
  - which returns an updated Adspec;
  - the updated version is then forwarded in Path messages sent downstream.

RSVP Messages - 2/2

- Path messages are sent with the same source and destination addresses as the data, so that they will be routed correctly through non-RSVP clouds.
- Resv messages are sent hop-by-hop; each RSVP-speaking node forwards a Resv message to the unicast address of a previous RSVP hop.

Soft State - 1/5

- RSVP takes a "soft state" approach to managing the reservation state in routers and hosts.
- RSVP soft state is created and periodically refreshed by Path and Resv messages.
- The state is deleted if no matching refresh messages arrive before the expiration of a "cleanup timeout" interval.
- State may also be deleted by an explicit "teardown" message.

Soft State - 2/5

- At the expiration of each "refresh timeout" period and after a state change, RSVP scans its state to build and forward Path and Resv refresh messages to succeeding hops.
- Path and Resv messages are idempotent.
- When a route changes, the next Path message will initialize the path state on the new route, and future Resv messages will establish reservation state there.

Soft State - 3/5

- The state on the now-unused segment of the route will time out.
- Whether a message is "new" or a "refresh" is determined separately at each node, depending upon the existence of state at that node.
- RSVP sends its messages as IP datagrams with no reliability enhancement.

Soft State - 4/5

- Periodic transmission of refresh messages by hosts and routers is expected to handle the occasional loss of an RSVP message.
- The network traffic control mechanism should be statically configured to grant some minimal bandwidth for RSVP messages to protect them from congestion losses.

Soft State - 5/5

- The state maintained by RSVP is dynamic; to change the set of senders Si or to change any QoS request, a host simply starts sending revised Path and/or Resv messages.
- The result will be an appropriate adjustment in the RSVP state in all nodes along the path; unused state will time out if it is not explicitly torn down.

Differentiated Services

Architecture: RFC 2475, Dec. 1998

DS Fields: RFC 2474, Dec. 1998

Architecture - 1/2

- Differentiated Services Architectural Model
  - Differentiated Services Domain
  - Differentiated Services Region
  - Traffic Classification and Conditioning
    - Classifiers
    - Traffic Profiles
    - Traffic Conditioners
  - Per-Hop Behaviors
  - Network Resource Allocation

Architecture - 2/2

- Per-Hop Behavior Specification Guidelines
- Interoperability with Non-Differentiated Services-Compliant Nodes
- Multicast Considerations
- Security and Tunneling Considerations
  - Theft and Denial of Service
  - IPsec and Tunneling Interactions
  - Auditing

Overview of Architecture - 1/5

- 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.

Overview of Architecture - 2/5

- Service differentiation is desired to accommodate heterogeneous application requirements and user expectations, and to permit differentiated pricing of Internet service.
- The 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.

Overview of Architecture - 3/5

- The 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.
- Per-hop behaviors are defined to permit a reasonably granular means of allocating buffer and bandwidth resources at each node among competing traffic streams.

Overview of Architecture - 4/5

- 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.

Overview of Architecture - 5/5

- 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.

Key Abbreviations

- DS
  - Differentiated Services
- PHB
  - Per-Hop-Behavior
- SLA
  - Service Level Agreement
- TCA
  - Traffic Conditioning Agreement

Terminology - 1/19

- 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

Terminology - 2/19

- 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.

Terminology - 3/19

- 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
    - the DS Fields standard,
    - the DS Architecture Informational RFC, and
    - other differentiated services documents; usually used in reference to a node or device.

Terminology - 4/19

- DS domain
  - DS-capable domain; a contiguous set of nodes which operate with a common set of service provisioning policies and PHB definitions.
- DS egress node
  - 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.

Terminology - 5/19

- 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 the RFC 2474.
  - The bits of the DSCP field encode the DS codepoint, while the remaining bits are currently unused.
- DS node
  - a DS-compliant node.

Terminology - 6/19

- 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.

Terminology - 7/19

- 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.

Terminology - 8/19

- 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.

Terminology - 9/19

- 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.

Terminology - 10/19

- 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.

Terminology - 11/19

- 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.

Terminology - 12/19

- 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.

Terminology - 13/19

- 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.

Terminology - 14/19

- 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.

Terminology - 15/19

- 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.

Terminology - 16/19

- 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. *(continues)*

Terminology - 17/19

- 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.

Terminology - 18/19

- 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.

Terminology - 19/19

- 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.

DS Architectural Model - 1/2

- Differentiated Services Domain
- Differentiated Services Region
- Traffic Classification and Conditioning
  - Classifiers
  - Traffic Profiles
  - Traffic Conditioners
- Per-Hop Behaviors
- Network Resource Allocation

DS Architectural Model - 2/2

- 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.

Differentiated Services Domain - 1/4

- 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.

Differentiated Services Domain - 2/4

- 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.
- 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).

Differentiated Services Domain - 3/4

- A DS domain normally consists of one or more networks under the same administration.
- 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.

Differentiated Services Domain - 4/4

- 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.

Differentiated Services Region - 1/2

- 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.

Differentiated Services Region - 2/2

- 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.

Traffic Classification and Conditioning - 1/4

- 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.

Traffic Classification and Conditioning - 2/4

- 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 Classification and Conditioning - 3/4

- 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.

Traffic Classification and Conditioning - 4/4

- 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 the DS architecture.

Classifiers - 1/2

- Packet classifiers select packets in a traffic stream based on the content of some portion of the packet header.
- Two types of classifiers are defines.
  - 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 - 2/2

- The classifier should authenticate the information which it uses to classify the packet.

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.
- Example:

  codepoint=X, use token-bucket r, b

  – r: rate
  – b: burst size

Traffic Conditioners - 1/4

- 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.

Traffic Conditioners - 2/4

- 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.

Traffic Conditioners - 3/4

```
                           +-------+
                           |       |-------------------+
                    +----->| Meter |                   |
                    |      |       |--+                |
                    |      +-------+  |                |
                    |                 V                V
              +------------+      +--------+      +---------+
              |            |      |        |      | Shaper/ |
packets =====>| Classifier |=====>| Marker |=====>| Dropper |=====>
              |            |      |        |      |         |
              +------------+      +--------+      +---------+
```

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).

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.

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.

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.
- A can be implemented as a special case of a shaper by setting the shaper buffer size to zero (or a few) packets.

Traffic Conditioners - 4/4

- 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.

Per-Hop Behaviors - 1/7

- 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

Per-Hop Behaviors - 2/7

- 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.

Per-Hop Behaviors - 3/7

- 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.

Per-Hop Behaviors - 4/7

- 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).

Per-Hop Behaviors - 5/7

- 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.
- PHBs are implemented in nodes by means of some buffer management and packet scheduling mechanisms.

Per-Hop Behaviors - 6/7

- PHBs are defined in terms of behavior characteristics relevant to service provisioning policies, and not in terms of particular implementation mechanisms.
- A PHB is selected at a node by a mapping of the DS codepoint in a received packet.

Per-Hop Behaviors - 7/7

- 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.

Network Resource Allocation - 1/3

- 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.

Network Resource Allocation - 2/3

- 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.
- The precise nature and implementation of the interaction between these components is outside the scope of DS architecture.

Network Resource Allocation - 3/3

- 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.

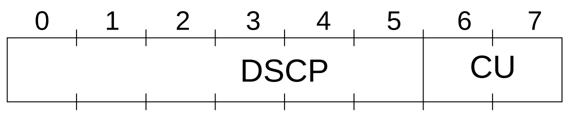
Differentiated Services Field Definition - 1/8

- A replacement header field, called the DS field, is defined, which is intended to supersede the existing definitions of
  - the IPv4 TOS octet [RFC791] and
  - the IPv6 Traffic Class octet.
- Six bits of the DS field are used as a codepoint (DSCP) to select the PHB a packet experiences at each node.

Differentiated Services Field Definition - 2/8

- A two-bit currently unused (CU) field is reserved for future usage.
- The value of the CU bits are ignored by differentiated services-compliant nodes when determining the per-hop behavior to apply to a received packet.

Differentiated Services Field Definition - 3/8



DSCP: Diffrentiated Services CodePoint
CU: Currently Unused

Differentiated Services Field Definition - 4/8

- DS-compliant nodes MUST select PHBs by matching against the entire 6-bit DSCP field, e.g., by treating the value of the field as a table index which is used to select a particular packet handling mechanism which has been implemented in that device.
- The value of the CU field MUST be ignored by PHB selection.
- The DSCP field is defined as an unstructured field to facilitate the definition of future per-hop behaviors.

Differentiated Services Field Definition - 5/8

- Packets received with an unrecognized codepoint SHOULD be forwarded as if they were marked for the Default behavior, and their codepoints should not be changed.
- Such packets MUST NOT cause the network node to malfunction.
- The structure of the DS field shown above is incompatible with the existing definition of the IPv4 TOS octet in [RFC791].

Differentiated Services Field Definition - 6/8

- The presumption is that DS domains protect themselves by deploying re-marking boundary nodes, as should networks using the RFC 791 Precedence designations.
- Nodes MAY rewrite the DS field as needed to provide a desired local or end-to-end service.
- Specifications of DS field translations at DS boundaries are the subject of service level agreements between providers and users.

Differentiated Services Field Definition - 7/8

- The DSCP field within the DS field is capable of conveying 64 distinct codepoints.
- The codepoint space is divided into three pools for the purpose of codepoint assignment and management:
  - a pool of 32 RECOMMENDED codepoints (Pool 1) to be assigned by Standards Action,
  - a pool of 16 codepoints (Pool 2) to be reserved for experimental or Local Use (EXP/LU), and
  - a pool of 16 codepoints (Pool 3) which are initially available for experimental or local use, but which should be preferentially utilized for standardized assignments if Pool 1 is ever exhausted.

Differentiated Services Field Definition - 8/8

```
Pool        Codepoint space          Assignment Policy
----        ---------------          -----------------
 1            xxxxx0                 Standards Action
 2            xxxx11                 EXP/LU
 3            xxxx01                 EXP/LU (*)
(*) may be utilized for future Standards Action allocations as
necessary
```