Inter-provider Quality of Service

White paper draft 0.13

September 20, 2005

[NOTE TO READERS: this is a work-in-progress. There are comments/asides included in square brackets that offer feedback, raise questions, or flag issues that need further work]

Table of Contents

Executive Summary

This document presents a straw-man proposal to enable the deployment of Inter-provider Quality of Service (QoS). We begin from the observation that QoS based on the Differentiated Services architecture [Blake98] is now widely deployed within the networks of single providers. This is especially the case for providers of Network-based VPNs (see, for example [RosRek05]). Some providers are now beginning to interconnect with each other via "QoS-enabled peerings" in an attempt to offer QoS that spans the networks of multiple providers. However, in the absence of appropriate standards and established procedures for management, trouble-shooting, monitoring, etc., such interconnections are likely to be challenging and labor-intensive. This document seeks to identify the key issues that service providers need to agree upon if inter-provider QoS is to be readily deployable.

This paper has two main goals:

- To identify standards that should be worked on to simplify deployment of inter-provider QoS
- To identify "best common practices" that, while not requiring standardization, could ease the deployment of inter-provider QoS if agreed to by a critical mass of providers

While there is plenty of debate about how many service classes need to be supported across multiple providers, it is widely agreed that some moderate number of classes should be commonly supported and consistently defined among providers. In this paper we take the approach of defining just a single additional service class (i.e., a single class which is offered as a service to customers, in addition to the best effort class). This discussion is offered as the simplest multi-class service offering, as a way of exposing the issues that must be addressed. The additional service class that is defined is intended to be suitable for real-time voice applications; and is intended to be appropriate for use both in a provider-provisioned VPN context and in the public Internet. We also note that in many cases providers may internally make use of an additional class of service that is restricted to network control traffic (such as routing protocol traffic and network management traffic).

The key issues that are addressed in this paper are:

Consistent Definitions of Metrics. To support QoS meaningfully across multiple providers, it is essential the metrics such as delay, jitter and loss are defined consistently.

Service Class Definition. The "real-time" service class is defined in terms of what the customer must do to receive the service (e.g. mark packets with a certain DSCP, conform

to a certain token bucket) and what the provider in turn commits to deliver (e.g. statistical bounds on loss, delay, availability).

Measurement, Monitoring and Reporting. Because of the multiple parties involved in the delivery of QoS, it is necessary to have defined methods for measurement of QoS, ways to monitor the performance of different network segments, and ways to report performance consistently among providers. We define such methods in this paper.

Routing. It may be necessary to route QoS-sensitive traffic to different providers or along different routes than those followed by best effort traffic. We define mechanisms that can be deployed to achieve these goals.

Provider Responsibilities. Finally, there may need to be some agreed-upon responsibilities and "best common practices" to which providers should agree. We propose a set of such practices with the potential to simplify deployment of inter-provider QoS among a large set of providers.

Suggested Objectives:

- This document is concerned with providing a universal definition of a small set of service classes, which will be transported between multiple providers from source to destination, and vice versa.
- This document is focused on the treatment that should be applied to two applications: a real time service, suitable for voice, and best effort.
- Although this document only outlines detailed criteria for the two aforementioned classes of service, its goal is to remain flexible so that additional (undefined) applications may be added to these existing service classes or to allow expansion for additional classes of service. We attempt to not limit the type and number of classes of service that may be used within an individual Service Provider, (although they may be subsumed into additional, Inter-Provider service classes in the future).
- In this document, not going to be concerned with Layer-2 (e.g.: Ethernet 802.1p, etc.) CoS. Although possible, it will be deferred to future consideration. This document will only focus on IP and MPLS CoS.

1 Introduction

[Placeholder text. Intro will be written once rest of document is complete. For now, using "things roughly agreed to" points as major points of consensus to guide drafting of white paper. Any text that violates these principles is something that should be flagged and we should discuss if there is consensus that change is appropriate. This will be deleted from the final draft.]

Things Roughly Agreed To (Version 1)

1. Need to aggregate customers into few classes at provider boundaries
2. Start with small number of classes supported by many providers (e.g. 2)

3. Common markings for the classes at provider boundaries considered desirable (IP & MPLS)
4. Providers at liberty to offer more than the minimum set of classes
5. Consistent metric set and their definitions, including common timescales, required
6. Means to measure & report said metrics need to be standardized
7. Allocation of metrics across multi-provider need to be worked out
8. Sharing of measurement data across providers is not a given
9. Near term routing needs can be addressed by today's BGP & application-layer routing
10. Exchange of QoS capability via routing (e.g. BGP) needs to be worked on
11. Robust recovery mechanism to handle failure of interconnect is required
12. Impairment budgeting/resolution needs to be worked on
13. Reference model needs to be agreed on

Things roughly agreed to v2

1. Common DSCP/EXP markings at provider boundaries considered desirable (IP & MPLS) associated with a defined service
2. Don't remark VPN customer DSCPs
3. Allow for IP, MPLS (a,b,c) or L2 interconnect
4. Ability to route different classes on different paths considered desirable
5. There exists a decent strawman BGP-based approach to (4).
6. Still need to define metrics (e.g., Subset of IPPM, 1541) including timescales (essential)
7. Bounds on the metrics for the "2nd" service should be defined (suitable for VOIP)
8. Don't require providers to open their networks to external measurement (but allow them to do so - level of monitoring depends on trust)
9. Loss, jitter, availability and delay are the key metrics
10. Simple allocation of delay budget across providers (e.g. 5ms + speed-oflight per provider) may be sufficient to get started
11. Roman's slide 3 is a good strawman for measurement agreements

2 Reference model & terminology

2.1 Definitions

Customer: The user of the services provided by a service provider. In the context of IPVPNs, a customer typically exists at multiple physical locations, all of which are under one administrative authority, but that connect to multiple VPN Service Providers. In the context of the Internet, the set of customers typically exist at multiple physical locations, each of which may be maintained by different different administrative authorities, and that connect to many different Internet Service Providers.

Provider: a single Internet and/or VPN Service Provider. In the context of this document, more than one Provider is required to deliver an end-to-end Quality of Service connection for the service class(es) defined herein.

P: Provider routers. A Backbone Router, within an Internet or VPN Service Provider(s) Network, that only attaches to PE's of the same Internet or VPN Service Provider.

PE: Provider Edge router. The router at the edge of a provider's network, usually facing towards a customer.

CE: Customer Edge router. The router at the edge of a customer's network, usually facing towards a provider.

CPI: Customer to Provider Interface. The physical link between a customer and a single Provider. This may also be referred to as a CE to PE or CE-PE link.

PPI: Provider to Provider Interface. A single, physical link between two, different Providers. This may also be referred to as a PE to PE or PE-PE link.

Interprovider Link: The link between two providers. Such a link typically interconnects a pair of ASBRs.

ASBR: Autonomous System Border Router. The router at the edge of an autonomous system (AS), facing towards another AS. ASBRs may be located at interprovider boundaries, or at AS boundaries that are within a single provider when a provider has chosen to divide his network in to several ASes.

Managed CE: A Customer Edge device that is configured and managed by the provider on behalf of the customer.

Unmanaged CE: A Customer Edge router that is only managed by the customer.

Option A (B, C): Methods for interconnection of MPLS VPNs across service provider (and AS) boundaries, defined in [3]

Measurement POP: A service provider's point of presence (POP) that contains equipment capable of responding to measurement probes from another location.

Trust Boundary: The line between two entities that do not fully trust each other. A CE-PE link is a typical example of a trust boundary because the provider does not trust the customer to configure his equipment correctly or to stay within his SLA parameters. Conversely an internal link inside a single provider's network is usually not a trust boundary.

Acronyms used herein include the following:

LSP Label-Switched Path

MPLS Multi-Protocol Label Switching

SLA	Service Level Agreement
DSCP	DiffServ Code Points
EF	Expedited Forwarding
EXP	Experimental (field) (e.g., in MPLS header)
VC	Virtual Circuit
BGP	Border Gateway Protocol
ATM PVCs	Asynchronous Transfer Mode Permanent Virtual Circuit
F/R DLCI's	Frame Relay Data Link Connection Identifier
PPI	Provider to Provider Interface
CPI	Client to Provider Interface
VoIP	Voice over IP
IPVPN	Internet Protocol Virtual Private Network
IP	Internet Protocol
PWE3	Pseudo-wire emulation edge to edge
GigE	Gigabit Ethernet

2.2 Scope

It is the intent of this document to develop solutions that are applicable for two major scenarios: the interconnection of ISPs, and the interconnection of VPN service providers. Because QoS deployment is much more well established in the VPN context than in the public Internet, we will use VPN provider interconnection as our primary focus, but the intent is to produce solutions that are applicable in the broader Internet context as well.

Within the VPN context, it is likely that many VPN providers will deliver a service based on RFC2547. This document will not restrict itself to 2547 VPNs - any IP VPN service should be supported - but we will address the specific QoS issues of interconnecting providers of 2547 VPNs, including the MPLS-based interconnection styles (referred to as options (b) and (c) in draft-ietf-l3vpn-rfc2547bis-03.txt).

2.3 Reference model

For simplicity, we consider first the single provider case depicted in Figure 1. In this model, as in [RFC2547], customer sites connect to the provider via a CE (customer edge) device, and the provider's routers that connect to customer sites are PE (provider edge) devices. The CE-PE link represents the typical boundary of trust between the provider and the customer. It may be possible to move the trust boundary to the CE if the provider manages the CE – we will consider this case below after treating the customer-managed CE case.

In the customer-managed CE model, it is the responsibility of the customer to ensure that the traffic that traverses the CE-PE link is "correctly" marked before it reaches the PE. "Correct" in this context simply means that the customer needs to ensure that packets are marked in a way that ensures they receive the service desired. For example, if the customer has subscribed to a "low latency" service and the SLA for this service dictates that packets must be marked "EF" to receive the service, then the customer must decide which of his packets are to receive the low latency service and mark them before they arrive at the PE. The selection of packets to receive the low latency service is thus entirely up to the policies of the customer.

The PE may enforce various aspects of the SLA, such as policing the amount of "EF" traffic received from a given customer. The details of such policing will be an aspect of the SLA definition; this topic is addressed in Section 3.

We note that the reference model places no restrictions on the mechanisms that are deployed by the provider within his core network. Services will be defined in Section 3 in terms of externally measurable performance parameters (e.g. loss, delay), with the mechanisms for achieving those performance targets left to the provider.

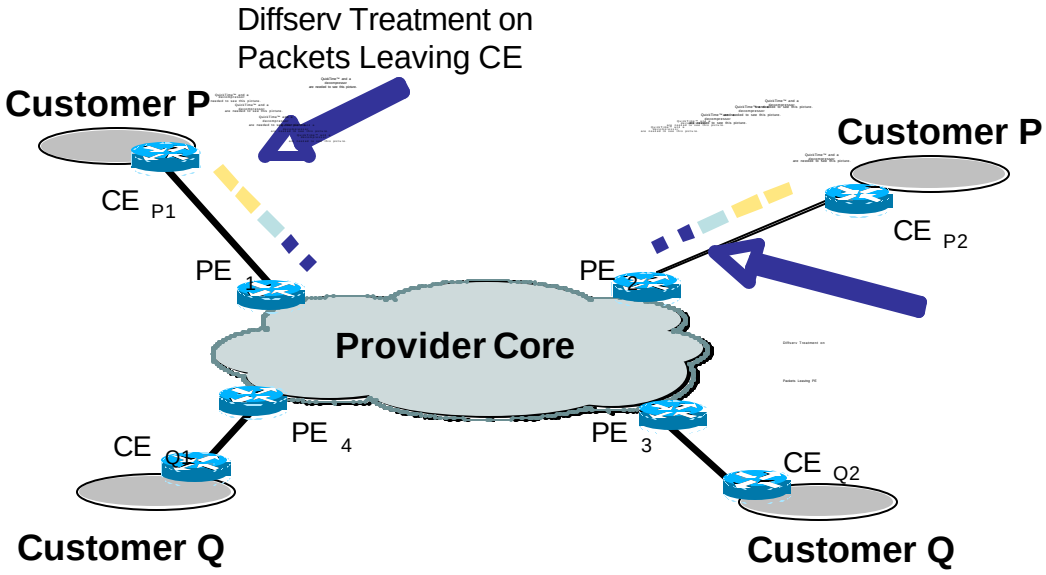


Figure 1. Basic Diffserv Model for Single Provider

Figure 2 illustrates a simple interprovider scenario. Its main difference from Figure 1 is that there are now two providers in the path between the two sites of each customer.

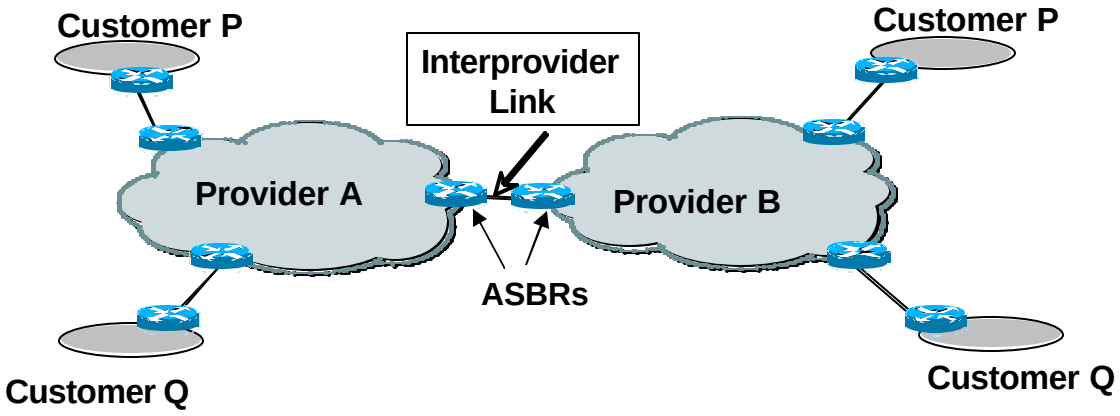


Figure 2. Simple Interprovider Topology

When we consider the problem of delivering a particular service (e.g. the "low latency" service) to customer P, several issues that were not present in the single provider case must be addressed, including:

- Packets must be "correctly" marked on the inter-provider link to obtain the desired service
- It may be desirable to carry that marking in a manner that avoids modification of the customer's data packets, e.g. in an extra header
- Providers A and B must each offer a service that, when concatenated with the service of the other provider, provides a useful end-to-end service to the customer (e.g., for a service with a fixed maximum delay, the allowable delay may need to be budgeted between multiple carriers).
- Monitoring the end-to-end performance experienced by the customer is now likely to involve both providers in the path.

Marking on interprovider links is the subject of the following section. Service definitions are discussed in Section 3. Measurement issues are discussed in Section 4.

It is desirable to support a wide range of interconnection methods. It should be possible to support a simple IP interconnect (which would include "option A" interconnection of RFC2547 VPNs) as well as MPLS interconnects of both option B and option C styles [3]. Interconnection using MPLS traffic engineered LSPs should also be possible. The encapsulation and style of interconnection used at the inter-provider boundary has consequences on the marking and policing requirements, discussed below.

2.3.1 Managed CE model

When the service provider manages the CE devices on behalf of the customer, it is possible to move the trust boundary to the CE. This means that the CE rather than the PE will be responsible for ensuring that the amount of traffic sent along the PE-CE link for

any service class does not exceed the SLA parameters for that service class. This may be achieved by policing or shaping of the customer traffic before sending it to the PE.

The managed CE case is more complicated when there are multiple providers as in Figure 2. If, for example, customer P purchases a managed service from provider A, who manages all of the customer's CEs, then the link between customer P and provider B still represents a trust boundary, while the link between customer P and provider A does not. In summary, the management of CEs by providers may or may not cause trust boundaries to be different than in the unmanaged CE case.

2.4 Marking

There is general agreement that customer packets should not be remarked (that is, have their DSCP values modified) as they transit the providers' networks. At the same time, it is often necessary for the provider to impose a QoS treatment on customer packets that differs from that which might be indicated by the customer's DSCP. A simple example arises when a customer has an SLA that allows him to send a limited amount of EF traffic into the network. If he sends more EF traffic than allowed, the provider may wish to "downgrade" the excess to best effort. However even if the packets are treated as best effort by the provider, the customer wishes to retain the EF marking for his own use when the packets arrive at his remote site. In the single provider environment, this capability is readily provided by encapsulating the customer's data with a header that exists only in the service provider network, e.g. an MPLS label header. This header is used to carry the Service provider's desired marking for the traffic.

When there are multiple providers in the path, as in Figure 2, the marking issue is slightly more complex. Packets need to be marked appropriately to receive the desired service from the provider on the receiving side of the link. (That is, packets from provider A need to arrive at the edge of provider B with an appropriate marking for the desired service.) In a pure IP (or option A) interconnect this leaves little choice other than remarking of the customer's headers. In options B and C, or when MPLS-TE is used across the inter-provider boundary, the MPLS EXP header may be used to carry the marking, thus leaving the customer header unchanged. It may also be possible to encode the marking in a layer-2 dependent way. For example, an 802.1q header may be used to carry the marking across the boundary, or multiple ATM VCs may be used, one per service, with provider A placing the packets on the appropriate VC to receive the desired service from provider B and *vice versa*.

2.5 Routing

In a network as simple as that shown in Figure 2, there are no real routing issues since there is only one path between any two customer sites. However it is clear that in a true multiprovider environment there may be many alternate paths between customer sites. The preferred path among providers is typically determined by BGP policies. However, when multiple classes of service exist, it may be desired to route some traffic preferentially via providers who support the enhanced QoS class(es) while best effort traffic takes the conventional route. This issue is addressed in detail in Section 5.

2.6 Measurement

Measurement is both important and challenging in the interprovider QoS context because of the relatively large number of providers in the path between two customer sites. In the single provider case, a customer can conduct performance measurement between CEs; if the performance targets are not met, it can be assumed that the problem lies with the provider (unless of course the customer has overbooked and thus congested the PE-CE links). Even in a network as simple as the one shown in Figure 2, there are now five possible locations of performance problems for a given site-site pair: the two CE-PE links, within the networks of each of the two providers, and the interprovider link.

In order to deal with troubleshooting and performance monitoring issues, measurement needs to be addressed as part of an Interprovider QoS solution. This topic is addressed in detail in Section 4.

3 Service Class Definition

3.1 Service Classes

This document is primarily concerned with the definition of a single, Expedited Forwarding service class, specifically for transport of Voice over IP (VoIP). This document also defines two secondary, non-real-time service classes, Best Effort and IP Control. The combination of these three service classes is considered integral to facilitate a basic Voice and Data service offering between disparate IP VPN carriers.

[Comment: An IP control has been added. Although it is reasonable to have such a class I think we are mixing services and classes here. If we want to include a discussion of an IP control class then maybe we could move it to a separate section so that we don't have to consider all the details of this additional class?]

3.1.1 Expedited Forwarding (EF) Service Class

The first service class defined is the Expedited Forwarding (EF) Service Class, which is to be used for the transmission of services that require low delay, low jitter and low loss between two, or more, disparate carriers. The EF Service Class makes use of the Expedited Forwarding Per-Hop Behavior (PHB) as defined in [RFC 3246], in addition to two, real-time, jitter sensitive, interactive (Class 0 and Class 1) services classes as defined in [Y.1541].

This document confines the use of the EF service class to be used specifically for the transport of signaling and media packets directly associated with the end-to-end transport of real-time, VoIP traffic. Although this EF service class is potentially suitable for transport of other real-time services, with similar latency, jitter and loss requirements as VoIP, their specification is beyond the scope of this document, but may be added to this service class in the future.

VoIP signaling packets are expected to be comprised of SIP, H.323 or MGCP packets, which are used to establish and tear-down media channels. VoIP media packets are confined to Real Time Transport [RTP] and Real Time Control Protocol [RTCP] packets, which are used for transmission of interactive, real-time speech and performance monitoring of the associated media channel, respectively.

Under normal operating conditions, the EF service class must carry the associated signaling and media packets for an individual VoIP call in both directions, (from source to destination, and vice versa). There is no requirement to provide a symmetric path for the bidirectional traffic flow between any given source and destination, which uses the EF Service Class.

3.1.2 Best Effort (BE) Service Class

The Best Effort Service class is the second service class defined in this memo. This is the "default" service class used primarily for transmission of all other IP traffic, which has not been explicitly identified and associated with another service class in this document. The Best Effort service class has no guarantee with respect to latency, jitter, or packet loss; however, carriers are expected to make generous efforts to provide for satisfactory delivery of packets in this service class to their destination.

The Best Effort service class may, at times, carry VoIP traffic, but this is expected to only be under exceptional circumstances.

- Control: IP Network Control traffic, (e.g.: BGP). This is not a service class transported edge-to-edge or end-to-end; but is necessary for reachability information to be exchanged between domains.

3.2 Network Layer Protocols

- Service classes defined for both IP and MPLS encapsulated traffic.
- The treatment applied to packets of a given service class is (largely) the same, regardless of encapsulation. Encapsulation is, for the most part, only relevant to DS network elements to rapidly classify packets for the sole purpose of applying an appropriate treatment.
- Physical & Link Layer media types are irrelevant, because all CoS discussed in this document will occur above Layer-2, (e.g.: MPLS or IP CoS).

3.3 Service Class Boundary Definition

- Physical Link between two, adjacent entities, (e.g.: Customer → Provider, or Provider-1 → Provider-2).
- In general, ingress to a network element denotes demarcation of service, a.k.a.: Trust Boundary.
- First "managed" element by the first SP, must enforce the terms defined for a given CoS.
- Client to Provider Interface (CPI)
- Restricted to single customer per CPI, (e.g.: point-to-point circuits).

- Otherwise, get into complications with hierarchical QoS.
- Future reference models may account for either multiple, different customers on single access circuit or same customer with multiple, different services on a single access circuit.
- In the context of the CE-PE link, in the "unmanaged CE" model, trust boundary is defined at the first ingress PE. On the other hand, with the "managed CE" model, the ingress CE may be defined as the trust boundary, in which case it must enforce the terms defined for a particular CoS.
- May use either IP or MPLS encapsulation.
- Provider to Provider Interface (PPI)
- Otherwise known as internal peering or internal, infrastructure links.
- Each PPI is expected to contain traffic from multiple customers.
- In the context of the ASBR-ASBR link, the receipt of traffic by a SP's first ingress ASBR.

[Comment: I think most of this (except the note on one customer per CPI) is covered by the reference model and terminology.]

3.4 CPI Treatments

[Comment: Under Re-marking it is stated as a requirement that only voice traffic may be carried in this service class. I think this is a too hard requirement. I would like to interpret the definition "a class suitable for voice" not to mean a class limited to voice. I guess that this limitation has been added to limit the packet size? The price we pay is that we might have to add an interprovider service class later to accommodate e.g. video conferencing or we might have to change the guidelines to accommodate these services later.

The IP Control class is given a separate queue. Again I don't really consider this part of the class suitable for voice.

Should we add the pipe model of rfc 3270 when considering egress classification of packets?

I think the discussion on scheduling needs new wording. The current text is trying to set a standard for the dimensioning of the link in reference to the three service classes that have been discussed and that the percentage values are really dimensioning criteria for the link. So this is really about provisioning of the link. Maybe an operational issue?]

- Marking
  - 'Standard' Inter-AS IP DSCP marking for VoIP [Telephony (EF), Signaling (CS5)] and Best Effort (0) are drawn from Section 2.3, Figure 3 of draft-ietf-tsvwg-diffserv-service-classes-00.txt.
  - Newly defined 'standard' Inter-AS MPLS EXP marking, for inner-most VPN MPLS label, of 5 for Telephony and Voice Signaling Traffic. Best-Effort must use MPLS EXP of 0, for inner-most VPN label.

  - Refer to "PPI Treatments" section for discussion on when out-of-contract, (from the standpoint of improperly marked), traffic will be dropped.
  - Providers are free to offer additional "local-use" IP or MPLS markers for other classes of service above and beyond standard Inter-AS markings, which only traverse their network. Recommended to use local-use ranges for IP and MPLS markers for this purpose.
- Re-Marking
  - Customer (or the SP on the customer's behalf, e.g.: managed CE model) may mark, or re-mark, Voice Signaling or Media traffic to the standard Inter-AS IP DSCP or MPLS EXP markings for Best-Effort (0) to allow its transmission to other SP's, albeit with lower/no service guarantees for Best-Effort. The reasons, mechanism(s) and best practices to do so are discussed in Section 5, (Routing).
  - Customer's traffic that is not Voice (Signaling or Media) must be marked, or re-marked, (by either the customer or the SP on the customer's behalf), to the standard, Inter-AS IP or MPLS EXP markings for Voice or Best-Effort if it will be transmitted to other providers. See Section 3.5, PPI Treatments, for more details.
- Policing
  - Policing shall be defined to mean traffic received in excess of the metered rate, (plus a negligible burst size), will be dropped.
  - Policing of individual service classes shall be metered/policed in accordance with the guidelines for the Telephony, Signaling and Best Effort classes from Section 2.3, Figure 4, of draft-ietf-tsvwg-diffserv-service-classes-00.txt, with the following exceptions/clarifications.
  - On ingress to the first SP's network, and/or the ingress CE in the managed CE model, the SP may perform policing on an individual service class. This should be performed for the Voice service class. This may be performed for the Best Effort service class.
  - Egress policing shall not be used on CPI links, (that is a function of the scheduler).
- Egress Queuing, (from egress SP's PE to egress CE)
  - SP will configure a minimum of 3 queues at the egress PE to the egress CE.
    - Queue 1: The first queue must be a Strict Priority queue that will carry the Telephony (IP DSCP EF or MPLS EXP 5) and Signaling (IP DSCP CS5 or MPLS EXP 5) packets.
    - Queue 2: The second queue may be a weighted-fair, or similar, queue that services packets marked with Best-Effort (IP DSCP 0 or MPLS EXP 0).
    - Queue 3: The third, and final, queue may be a weighted-fair, or better, queue that services packets marked with Control (IP DSCP CS6 or MPLS EXP 6) on the PE-CE link.
  - The reason there is a 'minimum of 3 queues' is the SP may offer additional service classes local to their network, which may require

additional queues above and beyond those of standard defined in this document.
    - SP will place packets into appropriate egress PE's queue, based solely on IP DSCP, or in the case of a native MPLS service (e.g.: CsC), the MPLS EXP value.
- Egress Scheduling, (from egress SP's PE to egress CE)
    - Egress Scheduler must be configured such that Queue 1, (Voice), can consume up to a maximum of 40% of the link's bandwidth. Queue 2 (Best Effort) must be configured to consume up to 58% of the link's bandwidth. Queue 3 (Control) must be configured to consume only 2% of the link's bandwidth.
    - The egress scheduler should allow any queue (Queue 1 through 3) to burst up to line-rate when there is idle capacity on the egress PE to CE link. (Note, due to hardware limitations this may not be feasible, in which case the egress SP should disclose this to the transmitting SP).
- Shaping must not be used on the CPI links for the Voice service class. It may be used for the Best Effort or Control service classes.
- WRED should not be used on the CPI links, (ingress or egress).

3.5 PPI Treatments

[Comment: We had a discussion about re-marking and dropping at one of the phone conferences earlier but I don't think that the current text on marking and re-marking reflects this?

Again the IP control traffic is treated as part of queuing and scheduling sections.

Actually, do we really need egress queuing and scheduling on PPI to be part of service definition? Isn't policing enough? If we have a operational issues section maybe this belongs there?

Again, should pipe model be mentioned regarding classification?]

- Marking
    - 'Standard' Inter-AS IP DSCP marking for VoIP [Telephony (EF), Signaling (CS5)] and Best Effort (0) are drawn from Section 2.3, Figure 3 of draft-ietf-tsvwg-diffserv-service-classes-00.txt.
    - Newly defined 'standard' Inter-AS MPLS EXP marking, for inner-most VPN MPLS label, of 5 for Telephony and Voice Signaling Traffic. Best-Effort must use MPLS EXP of 0, for inner-most VPN label. Top-most MPLS label must conform to the same values, but only for transmission across the PPI.
    - Transmitting SP to PPI: Traffic that does not conform to the aforementioned standard, Inter-AS IP or MPLS marking should be dropped by the transmitting SP at their egress ASBR, before transmission onto the PPI – this helps the transmitting SP to determine if traffic is

improperly marked and attempting to traverse the PPI. (This may be accomplished with a 0 pps rate meter/policer or ACL).
    - Receiving SP from the PPI: Traffic that does not conform to the aforementioned standard, Inter-AS IP or MPLS marking should be dropped by the receiving SP at their ingress ASBR, (before acceptance of traffic). (This may be accomplished with a 0 pps rate meter/policer or ACL).
- Re-marking:
    - On receipt of packets from the PPI, SP must not re-mark packets to conform to the standard, Inter-AS IP or MPLS EXP markings.
    - On receipt of packets from the PPI, SP may encapsulate packet, using either IP or MPLS, to conform to a 'local-use' MPLS EXP values within their Backbone, which are different from the standard Inter-AS MPLS EXP values.
- Policing
    - Policing shall be defined to mean traffic received in excess of the metered rate, (plus a negligible burst size), will be dropped.
    - Policing of individual service classes shall be metered/policed in accordance with the guidelines for the Telephony, Signaling and Best Effort classes from Section 2.3, Figure 4, of draft-ietf-tsvwg-diffserv-service-classes-00.txt, with the following exceptions/clarifications.
    - Transmitting SP to PPI: The transmitting SP should police Voice or Best-Effort traffic at their egress ASBR(s) if the receiving SP also has policers, to aid in capacity planning or diagnosis of non-conforming traffic.
    - Receiving SP from the PPI: The receiving SP should police Voice traffic at their ingress ASBR(s), which are receiving traffic from the PPI. The receiving SP may police Best-Effort traffic at their ingress ASBR(s), which are receiving traffic from the PPI.
        - See below section titled "Egress Scheduling" for values to use for ingress policing by the receiving SP.
- Egress Queuing, (from first SP's egress ASBR to second SP's ingress ASBR)
    - First SP will configure a minimum of 3 queues at the egress ASBR.
        - Queue 1: The first queue must be a strict priority queue that will carry the Telephony (IP DSCP EF or MPLS EXP 5) and Signaling (IP DSCP CS5 or MPLS EXP 5) packets.
        - Queue 2: The second queue may be a weighted-fair, or similar, queue that services packets marked with Best-Effort (IP DSCP 0 or MPLS EXP 0).
        - Queue 3: The third, and final, queue may be a weighted-fair, or better, queue that services packets marked with Control (IP DSCP CS6 or MPLS EXP 6) on the ASBR-ASBR link.
        - The reason there is a 'minimum of 3 queues' is the SP may offer additional service classes with another provider, which may require additional queues above and beyond those of standard defined in this document.

  - SP will place packets into appropriate egress PE's, or ASBR's, queue, based solely on IP DSCP, or in the case of a native MPLS service (e.g.: CsC), the MPLS EXP value.
- Egress Scheduling, (from first SP's egress ASBR to second SP's egress ASBR)
  - Egress Scheduler must be configured such that Queue 1, (Voice), can consume up to a maximum of 40% of the link's bandwidth. Queue 2 (Best Effort) must be configured to consume up to 58% of the link's bandwidth. Queue 3 (Control) must be configured to consume only 2% of the link's bandwidth.
  - The above constraints are derived from Y.1541 (?), whose recommendation is to constrain the volume of Best-Effort traffic in comparison to Voice traffic, particularly for low-bandwidth circuits, (specifically, T1 [1.544 Mbps] circuits), in order to meet strict, jitter SLA guarantees.
  - The egress scheduler must not allow any queue (Queue 1 through 3) to burst up to line-rate when there is idle capacity on the ASBR-ASBR link, otherwise it will be dropped by policers configured on the receiving SP's ASBR(s).
- Shaping must not be used on the PPI links for the Voice service class. Shaping may be used on the PPI links for the Best Effort or Control service classes.
- WRED should not be used on the PPI links, (ingress or egress).

3.6 Definitions of metrics

3.6.1 Initial Considerations

- Real-time (RT) or voice traffic class is characterized by three network performance metrics: one-way latency, one-way packet loss, and one-way jitter.
- Additional metrics can also be defined for this traffic class but their use is less frequent. Those are: availability, connectivity, throughput, and packet reordering.
- There is a wide spread practice of reporting the two-way metrics or one-way metrics derived from two-way measurements. However, the preference is for the point-to-point one-way metrics, as they reflect the most accurately the performance of the network. Two –way metrics should only be accepted on exception basis. We have to acknowledge that this is contrary to the current praxis.

3.6.2 One-way Metrics

[Comment: There is, I believe, an ongoing discussion on the issue if the jitter definition.]

- Basic Metrics
  - one-way delay

- one- way jitter
- packet loss

Note:

Basic metrics have well defined targets:

- *one-way delay (with target of < 150 ms (e2e) and < 70 msec backbone (G.711)) (msec)*
- *one- way jitter (with target < 35 ms (transport budget)) (msec)*
- *packet loss - (with target of 0.25% (e2e) (one-way)) (%)*

- Additional metrics
  - Throughput (bps)
  - availability – may be equivalent to packet loss – 0% packet loss 100% availability, or
  - availability/ connectivity - 5 9s with 0.9 sec downtime per day
  - packet reordering (no targets ?)

Conditions for metric delivery:

- All Performance guarantees are only for confirming packets/traffic
- ….

- Metrics are always defined by the relevant single instance of the metric measurement and the reported statistics of the metric. Single measurements are rarely reported and rarely stored in the network wide operationalized performance measurement systems. Single measurements are used and reported during the debugging or calibration process.

One –way delay (one-way latency) [OWD]

- The single instance of the one-way delay measurement is defined as the time the test (we assume here that we conduct active measurements with an injected test traffic) packet traverses the network segment(s) between two reference points.
  - The Metric is defined as a time from the time first bit of the packet is put on the wire at the source reference point to the time the last bit of the packet is received at the receiver reference point.
  - One-way latency is defined as a latency measured for packets traversing the network segment between the source reference point and the destination reference point
- The OWD metric is reported as a statistics (arithmetic mean) of several (specified) single measurements over a specified period of time
- The metric is defined for IP packets with RTP/UDP payload.
  - we are eliminating any special control protocols like ICMP
- End reference points of the measured network segment are synchronized to the external TOD
- In practice, on high capacity interfaces (OC-3 +), the interface insertion time could be neglected
- Defined should be overall test duration, packet separation, and packet IP QoS class.

- The size of the packet must be defined.
- The OWG is reported in msec; accurate to 1 msec, rounded up.
- The minimum reported one-way latency is 1 msec.
- Rejected from the calculation are:
    - All packets that are erroneous
    - Lost packets (infinite latency)
- [Reference RFC….]

One-way Jitter [OWJ]

- A single instance of the one-way jitter measurement is defined as a difference between the inter-packet gap at the source and the inter-packet gap at the destination for (two) consecutive packets in sequence separated by one or more other packets;
    - It is suggested that the number of the separation packets is 0. The number of separation packets (or a sampling function) has to be agreed to for the OWJ to have unequivocal meaning.
    - One-way jitter is defined as a jitter measured for packets traversing the network segment between the source reference point and the destination reference point.
- The OWJ metric is reported as a statistics (arithmetic mean) for a series of absolute values of individual measurements.
    - Absolute values eliminate negative jitter
- The metric is defined for IP packets with RTP/UDP payload.
    - we are eliminating any special control protocols like ICMP
- End reference points of the measured network segment are synchronized to the external TOD
- In practice, on high capacity interfaces (OC-3 +), the interface insertion time could be neglected
- Defined should be overall test duration and packet IP QoS class.
- The size of the packet must be defined.
- The OWJ metric is reported in msec; Accurate to 1 msec, rounded up.
- The minimum reported one-way jitter is 0 msec
- Rejected from the calculation are:
    - All packets that are erroneous
    - Lost packets (infinite latency)
- [Reference RFC]

Packet loss ratio (packet delivery ratio) [PLR]

- A single instance of packet loss measurement is defined as a record of the packet sent by the sender reference point at the destination reference point. The record is either 1 if the packet was received or 0 if the packet was not received or some additional conditions obtained that eliminated this packet from the packet count at the destination reference point (see this section for more specific on valid packets for the PLR metric).

    - Packet loss ratio is defined as a metric measured for packets traversing the network segment between the source reference point and the destination reference point
  - The PLR metric is reported as the percentage of sent packets from the sender reference point to received packets at the destination reference point.
    - The metric must specify
      - the number of packets that are send out in the test stream
      - the time interval packet are send out,
  - The metric is defined for IP packets with RTP/UDP payload.
    - we are eliminating any special control protocols like ICMP
    - the packets over which metrics are defined should be similar (in the stack composition) to the packets in the traffic stream, a performance of which is measured
  - End reference points of the measured network segment are synchronized to the external TOD
  - In practice, on high capacity interfaces (OC-3 +), the interface insertion time could be neglected
  - Defined should be overall test duration and packet IP QoS class.
  - The size of the packet must be defined.
  - Reported as % (0-100),
    - Could include the total number of sent packets, duration of the tests
    - The minimum reported packet los is 0%, the maximum reported packet loss is 100%
  - Rejected from the calculation are:
    - All packets that are erroneous
    - Packets received out of order
    - Packets received after some defined time limit.
- [Reference RFC]

3.6.3 Two-way metrics

[Comment: Should the two-way metrics still be in the white paper? I recall a discussion but not sure of the conclusion.]

- Two way metrics, in principle, are defined in the same way as one-way metrics.
  - *The definitions of the one-way metrics applies, then, to two way metrics.*
- The difference between two metrics is that the two-way metrics are defined over packets send and received by the same reference point, while the one-way metrics are defined over packets sent between two different reference points; one acting as a sender and one acting as a receiver. The two-way metrics have also defined two reference points a sender and a receiver reference points. However, in two-way metrics the receiver point acts only as a responder sending the packets back to the sender, and not keeping any statistics about the packets.

- In evaluating the two-way metrics one has to account for the possible asymmetry of the two-way route. The following approach is recommended:
    - The two-way latency metric should be bound by the estimated latency for the optimal route (2 x an air miles distance (~+ 30%)) between end points of the measured network segment.
    - There is no recommended way to set bounds on the two-way jitter and two-way packet loss to limit the effect of the asymmetric paths. The only possible way to bound those metrics is to report them always with the bound two-way latency metric.
- All metrics should be reported as point –to-point metrics, not as averages over the network 'cloud'.
- While two-way measurements do not need the synchronization of TOD of the end points to perform the measurements, the TOD synchronization is recommended/ required for the reporting purposes.

3.6.4 Additional metrics (TBD)

[Comment: Availability is mentioned as an additional parameter. I think it is a very important one that needs more discussion. Not sure if it is in this chapter or the SLA chapter or both though.]

- Throughput (bps)
- availability – may be equivalent to packet loss – 0% packet loss 100% availability, or
- availability/ connectivity - 5 9s with 0.9 sec downtime per day
- packet reordering (no targets ?)

Additional Issues

- Aggregation of budget for metric across multi-provider segments
  Other sections of the document
- Measurement and reporting methodologies
  Other segments of the document

3.7 SLA definition for the "low latency" class

[Comment: I don't think the delay values that are part of the ITU service classes are very valuable. Some service providers may write delay values related to geographic regions. Finding a general rule acceptable by all for how this is to be done will probably prove very cumbersome so it may end up to be part of every providers marketing strategy. The two rules mentioned as parts of the "budget allocation for planning issues" chapter (air distance*1.25*0.005 and air distance+5ms) are probably the best we can do.

I think we need to consider a concatenation rule for the delay variation to know what to recommend in the SLA.

Availability and service windows are mentioned here. Don't we need a more stringent SLA than to say that the values are guaranteed 95% of a month? The end goal should be a 24/7 definition but for the time being the customers need to be informed about when such maintenance windows might affect her traffic. Any opinions on this? This is probably an operational issue.]

Definitions of the following metrics are defined in the previous section. Methods and points at which they are observed are defined in Section 4, "Measurement".

The following are recommended maximum, or upper-bound, performance characteristics for IP or MPLS traffic that may traverse multiple providers. In large part, the following metrics were designed in consideration of very, low bandwidth links, (e.g.: T1). Therefore, these SLA guidelines may be tuned more aggressively, than the following recommendations, depending on the characteristics of provider's networks, access circuits, etc.

Y.1541, Section 5.3.6 of Y.1541 (5/2002) defines the following QoS classes relevant to this memo.

- Class 0: Real-time, jitter sensitive, high interaction (VoIP, VTC). This class applies to a service class whose traffic is "constrained [via] routing and distance".
- Class 1: Real-time, jitter sensitive, interactive (VoIP, VTC). This class applies to a service class whose traffic is "less constrained [via] routing and distance".
- Class 5: Traditional applications of default IP networks. This class applies to service class whose traffic may traverse any route/path.

- Class 0 and Class 1 apply to the Voice service class. More specifically, the Class 0 metrics apply to intra-continental Voice class between two service providers. Class 1 applies to the Voice service class extending on an inter-continental basis.
- Class 5 applies to the Best Effort service class.

Class 0:

- IPTD (IP Transfer Delay): 100 msec
- IPDV (IP Delay Variation): 50 msec
- IPLR (IP Loss Ratio): 1×10^{-3}
- IPER (IP Packet Error Ratio): 1×10^{-4}

Class 1:

- IPTD: 400 msec
- IPDV: 50 msec
- IPLR: 1×10^{-3}
- IPER: 1×10^{-4}

Class 5:

- All parameters are unspecified, hence no SLA is guaranteed for these packets.
- Y.1541 assumes that the above values are calculated on a 24 hours/7 days-per-week basis, unless specified otherwise. This document proposes that the above metrics are

determined over a period of 95% of one month, to allow for routine maintenance activities. This allows a service provider to have 3 periods of 3 hour maintenances per week. Over a course of a month this adds up to 36 hours attributable to maintenance activities.

- Recommend using a quantile-based (rather than mean) method for deriving, then calculating IPDV.

3.8 Impairment Budgets

3.8.1 Introduction

To support real-time traffic in multi-provider VPNs with the desired quality of service, the end-to-end impairment objectives for real-time-traffic should be met. The real-time Network QoS classes 0 and 1 of Y.1541 set these objectives. The topic of this section is the impairment allocation among multiple providers in order to meet those end-to-end objectives.

The guidance provided here is intended to accelerate the planning, deployment and management of networks and systems that can interoperate with a clear goal of supporting the end-to-end performance objectives detailed in Y.1541.

At the time of writing there are few examples of real-world deployments of multi-provider VPN with assured QoS, so there are no “common” or “best” practices. Discussions of algorithms to meet objectives within standards development bodies are ongoing. The algorithm suggested here was submitted for consideration to ITU-T. Before suggesting a particular algorithm we look at requirements and consider different general approaches.

3.8.2 Requirements

Any algorithm must be evaluated along with its probable implementation(s), which the following requirements reflect:

1) The algorithm should be
   a) Scalable - it should be able support paths between the many edges of every network provider.
   b) Robust – it should be widely applicable to the majority of situations including unusual topologies and distances, and recognize that capabilities of access and core networks are different (core network have multiple paths between points whereas access networks may not).
   c) Low overhead – the amount of extra traffic and extra infrastructure should be considered
   d) Timing appropriate to path selection needs – Business needs may dictate the need for frequent usage of allocations on multi-second, multi-month or indefinite sessions, starting immediately or at some time in the future.

e) As simple as possible but no simpler
f) Secure – considering
   i) Access Control
   ii) Authentication
   iii) Non-repudiation
   iv) Data Confidentiality
   v) Communication Security
   vi) Data Integrity
   vii) Availability
   viii) Privacy
g) Resistant to gaming – providers which don't meet expected objectives must be detectable.

2) Time sensitivity of solution
   a) The evolving nature of requirements and technology are recognized. Consideration of solutions should target particular deployment timeframes and evolving technology trends.
3) Consideration of how SPs handle cases where the aggregated impairments exceed those specified for a Network QoS Class

Some algorithms will, by their very nature support additional capabilities that are not seen as current requirements. For example, a provider may offer a menu of impairment capabilities between edges based upon offered financial cost. It is recognized that the evaluation of solutions may change if requirements change.

To help describe the various approaches we first define two terms as used in this paper.
Apportionment Method of portioning a performance impairment objective among segments
Allocation Formulaic division or assignment of a performance impairment objective among segments

3.8.3 Consideration of Approaches

[Comment: I like the addition of the accumulation approach. It would be even better if we could find requirements related to geography but this is easier said than done.]

Generally the approaches that could be taken in allocating total impairment targets among network segments can be characterized by the amount of information shared among segments. Each approach has their pros and cons, we describe them here.

For all approaches, a "top-down" or "bottoms-up" method could be applied. That is, percentages of the aggregated target (top-down) or fixed/negotiated values for impairments (bottoms-up) may be allocated for each segment. A hybrid of these methods, with percentages for some segments and fixed/negotiated values for others could also be used.

For some approaches, transit segment distances are required to estimate distance dependence metrics such as mean delay. Ground level distance between any two (User) points may be readily estimated despite the traffic's signal being carried over varying altitude, the non-spherical shape of the earth, etc. Distance-inefficient routing over multiple segments may result in traffic traveling over a significantly longer distance than expected between two User points. The approaches to accounting for these inefficiencies can also be characterized by the amount of information shared among segments. Selection of the quantization of distance e.g. kilometers, metro, regional, continental and international is independent in approaches where awareness of distance is required.

Regardless of the approach, there is no guarantee that the end-to-end objectives will be met.

The long term objective is expected to be a signaled approach, however, near-term, some simpler approach evolving to a more capable signaled approach may be recommended. In which case the recommendation should include an evolution path.

[replace the following tables with text, describe the major approaches with their pros and cons and other considerations]

Approach	Description	Information required at each segment	Pros	Cons
Static (simplest/least flexible) - no information is required to be shared among segments	A fixed number of segments is assumed Impairment allocation is formulaic among User, Access, Transit, and Peering segments	Information required is a) type of link, b) traffic service class and, c) transit distance	No information is required to be shared among segments. Access providers may re-allocate among their User, Access and Transit segments	May be over-engineered when number of segments is less that the number assumed Paths having more that the assumed number of segments are not covered Negotiation not supported.
Pseudo-static - some information is required to be shared among segments	The exact number of transit providers is determined Impairment allocation is formulaic among User, Access, Transit, and Peering segments	Information required is a) type of link, b) traffic service class and, c) transit distance d) destination address e) BGP tables	Impairment allocation may be efficient and scalable.	Signaling among providers is required to determine the number of transit providers in each traffic path e.g. from BGP number of AS's Negotiation not supported

Signaled (least simple/most flexible) - some information is required to be shared among segments and possibly with Users	The exact number and sub-type of all segments may be known e.g. if User segment is wireless or wireline Impairment apportionment may be negotiated among segments and with Users	Information required is a) type of link, b) traffic service class c) destination address d) BGP tables, or other means to determine path or paths at the operator-level, e) Network edge-edge performance information Additional information that may be required includes f) transit distance	Negotiation is supported allowing highly flexible apportionment among segments. No predefined allocations are required. Transit distance may not be required Able to address cases where the objective can not be met by consulting user for relaxed objective Consistent with proposed direction of methods automated by QoS Signaling (e.g. RSVP/NSIS).	Signaling among providers is required to negotiate the impairment apportionment for each segment. Signaling may be required to negotiate with User when the requested objective cannot be met Performance and routing information must be exchanged among providers to determine the identities of transit providers in each traffic path (e.g. from BGP number of AS's) and their performance. However, there are alternative ways to determine path, and many providers publish performance info in real-time.

Table 1 – Summary of performance impairment apportionment approaches

A fourth approach which does not start the apportionment process with target impairments per provider as a basis is called here the "Accumulation" approach, it is summarized in Table 2 below:

Approach	Description	Information required at each segment	Pros	Cons
Accumulation - some information is required to be shared among segments	The path through various network operator domains is determined. Impairment levels and other parameters may be solicited for various network segments, combined and compared with Desired Objectives. If not met, then Path or User negotiation takes place, or the request is rejected.	Information required is a) traffic service class, b) destination address (always known), c) BGP tables, or other means to determine path at the operator-level, d) Network edge-edge performance	No allocations required, so no process to achieve agreements Impairment accumulation is simple and scalable. No distance and route-to-air conversion factors required. Negotiation is supported. Consistent with future methods automated by QoS Signaling (RSVP/NSIS).	Performance and routing information must be exchanged among providers to determine the identities of transit providers in each traffic path (e.g. from BGP number of AS's) and their performance. However, there are alternative ways to determine path, and many providers publish performance info in real-time. Cannot guarantee that objectives will be met (true for all approaches to some extent).

Table 2 - Approach to Impairment Apportionment based on Accumulation

Compared to networks and systems that are circuit-based, those based on IP pose distinctly different challenges for planning and achieving the end-to-end performance levels necessary to adequately support the wide array of user applications (voice, data, fax, video, etc). The fundamental quality requirements for these applications are well understood and have not changed as perceived by the user; what has changed is the technology (and associated impairments) in the layers below these applications. The very nature of statistically multiplexed IP-based networks makes balancing capital efficiency with end-to-end performance across multiple network operators a very major challenge for applications with stringent performance requirements.

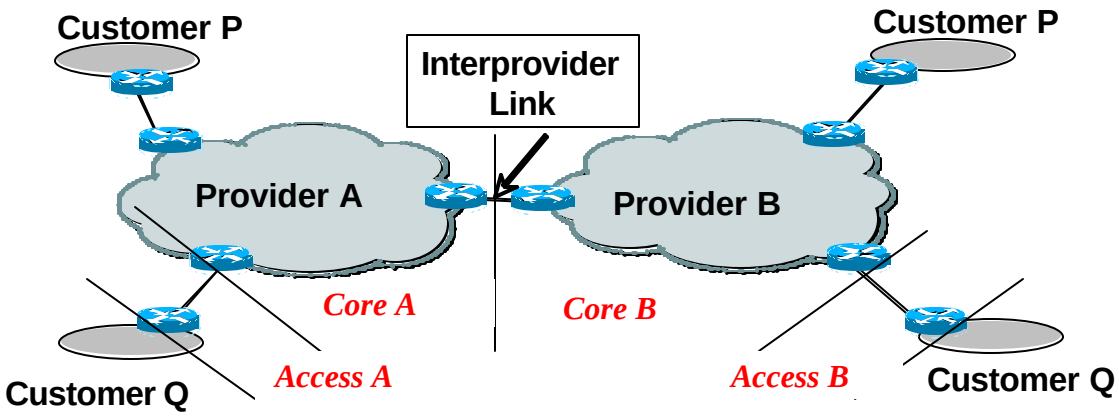
Section 3.7 outlines end to end targets for the classes being considered in this paper. These end to end targets are valuable to aid the definition of the service experience of an end user, but are of a lesser value in themselves to network planners.

Where an end to end service is provided across a single provider's network, the service planner can singly apportion impairments in various parts of the providers network for each service connection offered. The network planner has the full visibility of all network

components contributing to the overall outcome and can plan the resultant service in the manner that best fits the desired technical and commercial outcome.

Where an end to end connection spans multiple provider's networks this end to end visibility is no longer so readily available. Network planners need to understand the boundaries to work within that will result in a high probability that an acceptable outcome can be achieved across any reasonable combination of providers that may be required to collectively provide an end to end service.

The reference model used is as follows [this needs to be redrawn to look a little better]



For the purposes of impairment allocation, the edge of the providers core is the midpoint between ASBR's. The interconnecting link dimensioning may need to larger than might otherwise be required to ensure its contribution to the core allocation of the interconnecting providers allocations is not excessive.

3.8.4 Access network

The access network is that part of an end to end connection from the customer's side of the CE router to the first PE router.

For many network, the access network is the network domain where bandwidth per user is the most cost sensitive. Total bandwidth is therefore limited and in many cases may be a T1 or E1 link or less.

The ability for any access provider to dynamically change the size of the access link to compensate for higher consumption of the impairment budget in a core network for a given connection is very limited and in most case not operationally or commercially practical.

Long low capacity links using copper or radio based technologies are subject to higher interference than high speed optical core links. AS a consequence a significant proportion

of interference related impairment such as packet loss needs to be allocated to this part of a network to optimise the price performance of the overall outcome.

3.8.5 Number of providers

The number of providers in any end to end connection will vary based on both technical and commercial considerations.

It is assumed in this paper that for any national (trans United States or Australia) or regional (e.g. Western Europe or Eastern Asia) service, use of no more than three concatenated core network providers would be a reasonable maximum. The end to end budgets should have a high probability of being met if all providers consume their maximum impairment budget allocation.

In practice some negotiation or signalling of impairments between operators may be employed to ensure the end to end budget will be met or exceeded for any individual connection, but a single provider's network planners cannot rely on this.

For international connections or those between major global regions of North America, Asia Pacific, Central Asia and Europe it is assumed the end to end impairment targets will be those of ITU-T 1541 Class 1.

3.8.6 Budget allocation for planning purposes

[Comment: Is this only for planning purposes or are these values to be considered as requirements for service providers claiming to offer a service class suitable for voice? I think it could be used for both.

We are sort of stuck on the jitter concatenation rule. Or is this just not documented? Should we send in proposals?]

The approach allocates a significant proportion of the impairment budget to each access segment, with each core segment having a lesser fixed budget. The approach also allocates a fixed budget for core network segments, irrespective of the number of core network segments in any resulting services. This core network segment budget can be concatenated within bounds to create end to end services that have a high probability of still being within the overall end to end performance class targets.

An additional allowance for propagation delay for long network segments is also provided. Core network segments only need to have knowledge of the distance between their edges when the total distance between the edges of any core network segment exceeds an air path distance of 1200km.

For each core network segment, a combined budget is to be allocated for the numerical sum of IPTD and IPDV. In core network segments, the IPTD is largely determined by propagation times whereas IPDV is typically incurred as a result of network element processing and packet forwarding delays. The operator of any core network segment would be at liberty to allocate the total budget to either IPDV or IPTD or any mix of both, provided the numerical sum of IPTD and IPDV did not exceed the allocated budget

value. This poses some risk that the overall outcome may exceed targets if the entire budget is allocated to one metric, but real core network segments will almost always result in a mix of IPTD and IPDV.

The recommended allocations for the service class discussed is as follows;

	Access segment			Core network segment (<1200km)	
ITU-T Y1541 class	IPTD (200 byte packets)	IPDV (99.9 percentile - minimum)	IPLR	IPTD + IPDV (99.9 percentile – minimum)	IPLR
Class 0 and 1	<25ms	<15ms	<4x 10-4	<10ms	<1x10-5

Notes

1. Providers can trade off one way delay and delay variation in the core domain. The impairment allocation assumes the worst case, where the budget is predominantly one way delay. Where the budget used has a higher delay variation component the overall end to end outcome may well be more favourable due to the non-additive nature of delay variation.
2. This threshold allows for xDSL access technologies. Links using ADSL for example will attempt to maintain a minimum bit error rate of $1x10^{-7}$. A single error burst can result in the loss of a DSL frame or cell, and therefore loss of a packet.
3. For access segments with a peak data rate of under 2 Mb/s that are also used to carry best efforts traffic on the same access link as the low latency class traffic then packet fragmentation techniques needs to employed to enable the delay variation target to be achieved. This is to avoid a low latency class packet getting "stuck" behind a large best effort packet.

Where a single core network segment is greater than 1200km from edge to edge, an additional allowance for propagation delay is allocated. For this the following formula would apply;

Additional IPTD (ms) = (total segment air path distance in km -1200) x 1.25 x 0.005

The additional IPTD budget should be rounded up to the nearest integer number of milliseconds.

This approach requires lowest latency services (ITU Y1541 Class 0) to have no more than three core network segment providers. Typically no more than this would be used in any "national" or regional connection to achieve lowest latency performance. Inter continental services could only met ITU Y1541 Class 1 performance (IPTD relaxed to 400ms end to end) under this approach unless network segment providers negotiated

lower budgets for a service. For these longer path length services, the number of core network segment operators can be greater than three.

3.8.7 Application of impairment budgets

If we now consider worst case scenarios that may result. These occur when all participants in an end to end connection use their maximum impairment allocations. This situation will be rare in actual networks and real network elements cannot be that precisely configured.

We will assume the total air path distance is 4000km (e.g. Trans U.S.A.) and there are 3 core operators involved in the end to end connection.

3.8.7.1 Case 1: 100% of the core budget is one way delay

	Link air path distance	IPTD budget	Additional IPTD for Distance	Core budget used for IPTD	Total IPTD	IPDV (inc balance of core budget)	IPLR
Access provider 1		25ms			25ms	15ms	4×10^{-4}
Core provider A	300km		0	10ms	10ms	0	1×10^{-5}
Core provider B	3000km		12 ms	10ms	22ms	0	1×10^{-5}
Core provider C	700km		0	10ms	10ms	0	1×10^{-5}
Access provider 2		25ms			25ms	15ms	4×10^{-4}
Total CE to CE	**4000km**				**92ms**	**30ms**	**8.3×10^{-4}**

In this case the end to end one way delay exceeds the total end to end budget by 10ms, but the delay variation is well under target. For a voice service this would still result in a very satisfactory outcome.

In practice this case would never occur as some of the "non distance" related core budget would always be delay variation.

3.8.7.2 Case 2: 100% of non-geographic core budget delay variation.

	Link air path distance	IPDT budget	Additional IPDT for Distance	Core budget used for	Total IPDT	IPDV (inc balance of core	IPLR

				IPDT		budget)	
Access provider 1		25ms			25ms	15ms	4×10^{-4}
Core provider A	300km		0	0	0ms	10	1×10^{-5}
Core provider B	3000km		12 ms	0	12ms	10	1×10^{-5}
Core provider C	700km		0	0	0ms	10	1×10^{-5}
Access provider 2		25ms			25ms	15	4×10^{-4}
Total CE to CE	**4000km**				**62ms**	**60ms***	**8.3×10^{-4}**

* The total for IPDV in the table is shown as an additive sum. But concatenated IPDV is not additive. The net effect is that even in this extreme case it is likely the actual end to end IPDV would be within the 50ms IPDV target of Class 0.

3.8.7.3 Case 3: Transcontinental service 5 providers, ITU-T class 1

In this case, each core provider has elected to use their core network segment budget in different ways. The example highlights the impact of making these choice independently of other providers where more than three core network segments are used.

	Link air path distance	IPDT budget	Additional IPDT for Distance	Core budget used for IPDT	Total IPDT	IPDV (inc balance of core budget)	IPLR
Access provider 1		25ms			25ms	15 ms	4×10^{-4}
Core provider A	300km		0 ms	4 ms	10ms	6 ms	1×10^{-5}
Core provider B	3000km		12 ms	6 ms	22ms	4 ms	1×10^{-5}
Core provider C	10,000km		55 ms	5 ms	60 ms	5 ms	1×10^{-5}
Core provider D	2,000Km		5 ms	6 ms	11 ms	4 ms	1×10^{-5}
Core provider E	400Km		0 ms	3 ms	3 ms	7 ms	1×10^{-5}
Access provider 2		25ms		0 ms	25ms	15 ms	4×10^{-4}
Total CE to CE	**17,000km**				**156 ms**	**56 ms***	**8.5×10^{-4}**

* The total for IPDV in the table is shown as an additive sum. But concatenated IPDV is not additive. The net effect is that even in this extreme case it is likely the actual end to end IPDV would be within the 50ms IPDV target of Class 1.

Both case 2 and 3 assume worst case Delay variation based on this metric being additive. Because delay variation is not additive, the actual figure will be better than outlined. There are a variety of proposed methods [include references] for combining delay variation values.

4 QoS Measurement

[Comment: There are two-way measurements mentioned. Are these needed? I seem to recall a discussion on this where we concluded that one way was enough, is this right?]

The section should specify best practices to implement measurements of the network performance; to the extend that all SPs would agree that implementing such measurements is feasible and obtaining results from such measurements meaningful.

4.1 QoS Measurement Requirements

[Comment: CE-CE measurements and PE-PE measurements are mentioned. What about measurements between PE and ASBR and ASBR to ASBR (for transit segments). Maybe this is not a specific requirement since it is within a domain?

Do we need to define measurement point types in order to make publication and presentation of results simpler? E.g. Measurements performed directly from routers compared to measurements made from dedicated equipment connected to routers. Connection point such as PE, P or ASBR with reference making it possible to relate to traffic path? How about accuracy of measurements?]

- The measurement methodology, protocol and reporting must be capable of estimating at least the following QoS metrics defined in section 3.6 of packets transmitted between specified measurement points
    - One-way Packet Delay (OWPD)
    - Two-way Packet Delay (TWPD)
    - One Way Packet loss (OWPL)
    - Two Way Packet loss (TWPL)
    - Packet delay variation (PDV), sometimes also called "jitter"
    - Connectivity – Uni/Bi-directional, Instantaneous and Interval
- Periodic and on-demand measurements shall be supported
- The measurement probe packets must traverse as much of the same path and QoS mechanisms in routers as user packets having the same QoS service class.
-
- The measurement method and protocol
    - shall be capable of measuring QoS from a PE in one provider to a PE in another provider
    - should be capable of measuring QoS from CE to CE
    - The measurement method and protocol must provide means to limit and detect attempts to tamper with or alter the QoS metric estimates

- The measurement methodology shall not require that providers provide access to measurement points nor exchange measurement data. However, the protocols should support access to measurement points or measurement data between consenting providers for authorized requestors.
- SPs should agree to the ToD accuracy level (e.g., NTP stratum level of the measurement point)

4.2 QoS Measurement Methodologies

- Ideally, the measurement methodology would be common among providers; however, this may not be practical in the near to mid-term since at least the following measurement methodologies are either in use or are a possibility
- The measurement probe packets source and sink location may be:
  - Deployment of probes connected to specific ports on routers of the service(s) for which QoS parameters are to be implemented
  - Usage of protocols embedded within routers that provide the service
- The placement of measurement points may be different in provider networks
  - Each CE or a subset of CEs
  - Each PE router or a set of PE routers
  - Each P router or a set of P routers
- The measurements may be reported as
  - Point to point measurements between two measurement points
  - A matrix of point to point measurements between specified measurement points
  - An average between one or more specified set(s) of measurement points (An SP should avoid reporting measurements averaged over measurement points located in widely different geographic areas)

- Measurement traffic should not affect the production traffic
- If applicable, SPs should agree on the level of the test traffic (BW), either terminating in a measurement point or transiting their network
- SPs should agree the details of the measurements for reach type of measurement; in this, the recommendations of section 3 may be followed or the SP may choose to use its own methodology but then it should disclose it to the interested parties. The methodology should define the type of measurement packets (size, protocol); the frequency of tests, the frequency of packets in tests series; the interpretation of test results, the aggregation method; if active method is used for measurements.
- SPs should clearly publish how their measurement points for the measured segment.attach to their network with respect to the peering service provider's entry to that SP's network
- Measurement methodology should specify how the errors in measurements are treated, how results are processed in terms of any statistical treatment of data.

- The measurements may have at least the following
  - On-demand based upon operator or system command

  - Periodic measurements based upon pre-configured parameters that specify at least the number of probe packets over a measurement interval
- The measurement of QoS across multiple provider networks could use at least the following methods
  - Usage of a common measurement protocol with either probe points and/or embedded router function in each provider network along the user packet path
  - Usage of different measurement protocols across each provider network as measured between edge routers connecting these providers with per service class interface statistics used to estimate the concatenated end-to-end performance
- In order to support diagnostics and SLA conformance tracking, a provider must retain QoS measurement data for some agreed to period

4.3 Candidate QoS Measurement Protocols

- Ideally, the measurement protocol would be common among providers; however, this may not be practical in the near to mid-term since at least the following measurement protocols have been defined and are in use
    - ICMP-based PING measurement of TWPD, TWPL, and Instantaneous Bi-Directional Connectivity has historically been used by a number of providers in QoS-oriented SLAs
    - Vendor proprietary measurement protocols have been developed and used by some providers
    - IETF IPPM WG has been defining and some vendors have implemented a subset of the recommended measurements [Cite IETF I-D]
    - RTCP XR has been defined and can measure TWPD, OWPL, and PDV and has been used by some providers
    - TWAMP simplified (IPPM WG protocol)
  - Active vs. Passive
  - Timescales for probing, averaging, and reporting

4.4 Reporting of Measurement results

SPs shall agree on the reporting methods. At the minimum there should be the agreed upon process for exchange of hard copies of the performance results.

It is highly recommended that SPs agree on the electronic exchange of results via WEB site or periodic reports using basic content to be specified in this document.

The frequency of reports and details must also be agreed upon between SPs. It is proposed that reports shall be available at least daily. Note that the level of service offered (e.g., Availability or MTTR) may dictate more frequent reporting (e.g., hours or even minutes).

Instantaneous results should not be reported (single measurement). Only aggregated data should be reported. Aggregated data should be available at different aggregated levels (by the fraction of an hour, by hour, daily, monthly – depending on the report)

Basic distributional statistics of aggregates (mean, median, quantiles, number of measurements)

The report should include at the minimum:

- Date,
- Time,
- Location of end points
- Measurement/report period
- Measurement type
- Measurement statistics

4.5 QoS Measurement Security Considerations

- Reporting of measurement data (for customer, potential customer and troubleshooting) shall be subject to a standardized, secure authorization protocol
- Access to a measurement protocol probe or embedded router measurement point shall be subject to a standardized, secure authorization protocol
- The SP may open the measurement points in his/her network for outside access but with the controlled access (authentication).

5 Routing

While existing routing may be sufficient in some inter-provider QoS deployment scenarios, it may also be desirable to select among multiple interdomain paths based on the QoS requirements of different classes of traffic. That is, there may be cases in which the current route selection capabilities of BGP, which yield only a single best path for a given prefix, may not be sufficient. This section proposes an approach to extending BGP to address such cases.

Extending BGP to support QoS-aware routing inherently implies increasing the amount of information carried in BGP. This could have some implications for the convergence and scaling of BGP, at least in principle. Moreover, in order to maximize the stability of inter-domain routing in the Internet, it is highly desirable that the QoS-related information that is to be advertised into BGP be stable (in terms of not changing rapidly over time). These issues should be taken into consideration if BGP is extended to carry QoS-aware information.

Currently interdomain BGP peering is limited in its ability to distinguish NLRI ("prefixes") associated with different services (e.g. different QoS classes). The proposed approach to address this issue is to provide BGP with the means to mark address families (AFIs, SAFIs) and prefixes via a simple, opaque (to BGP) marking, to associate a them with a "service context" (e.g. QOS class). The approach draws on and extends the current work on "multisession BGP" described below.

The background for this work is the ever-changing environment around the destination tables for BGP [I-D.ietf-idr-bgp4] updates. Originally BGP was targeted at the global IPv4 unicast table. It was later extended with the Multiprotocol Extension[RFC2858] which allowed addressing different address families based on known address family and subsequent address family identifiers (AFIs and SAFIs) so that, for example, IPv4 multicast reverse path information could be propagated through BGP.

The Multiprotocol Extension enabled the use of multiple routing tables that are distinguished by their usage (e.g. IPv4, IPv6, VPN-IPv4, Multicast), but left little available to extending the original concept with other types of table separation. The type of separation that is desired for interprovider QoS is the ability to mark the existing update messages in a way that identifies the service contexts without having to force the use of a two level afi/safi hierarchy. It is important that any changes retain backward compatibility with existing BGP extensions, such as Route Refresh[RFC2918] etc.

There are other additional features that are needed to build a interdomain system for service separation that can enable revenue generating service level agreements. They include: BGP peering session separation, passing of redundant or backup routes, faster failure notification propagation and the ability to have 'service topologies' or network overlays and pass 'context' information within the new hierarchy. These are covered in later sections.

5.1 Current BGP Capabilities

BGP is good at passing end-to-end routing reachability between two peers. There are no additional semantics that the protocol is aware of that are carried in the update messages. All additional semantics attached to a prefix are opaque to the protocol (e.g. extended communities) and have local semantics. Unfortunately, BGP is not a suitable protocol for passing rapidly changing path characteristics (delay, jitter, etc) as the protocol is based on a distance vector architecture and not one that floods data or has full network topology awareness.

As noted above, BGP is also capable of carrying multiple classes of routing information through its AFI/SAFI hierarchy. QOS class or service context could be considered as a class of routes and BGP could simply announce reachability and service/QOS classes would be passed along in an opaque manner. If, as this paper proposes, there is a very small, bounded number of classes that are infrequently changing, this problem should be tractable. There are a few more problems that need to be solved with respect to the BGP protocol architecture before things would work perfectly. BGP has no way to carry multiple routes to the same destination. The protocol is based on "implicit withdraw" semantics. This means that every new announcement of a prefix causes any other announcement of the same prefix to be "withdrawn" or no longer reachable. Thus, announcing a prefix multiple times (e.g. once per QOS class) may not work well.

Also, BGP in most current implementations is based upon multiplexing all AFI/SAFI onto one BGP peering session, which implies shared fate in the state of the peering session. An error in one AFI/SAFI update message causes all prefixes in all AFI/SAFIs to be purged. Due to this multiplexing, it is also impossible to prioritize the convergence of the prefixes associated with one service, AFI or SAFI upon reception of a new update. All are treated equally in a "first in, first converged" manner.

5.2 Solution Assumptions

There are several options for a solution. We could define a new AFI/SAFI for each QOS class, have a distinct session for each service, agree upon or exchange all QOS markings via negotiation as some examples.

A few assumptions are in order to bound the problem and find a solution. It may be desirable to decouple the markings used for packet forwarding from the QoS class. This allows one provider to change their markings as they wish and to use different markings than their peer domain for greater flexibility in service offerings. Thus, only the link between the two domains would need to be administratively agreed upon. The solution set should allow for both multiplexing of services on one link as well as the use of logical links across which only one service type traverses. Last, it is assumed that the least disruptive change to the existing BGP protocol and protocol packet format would be best for ease of backwards compatibility, development and deployment.

An operator also may want to build specific service topologies within their domain. This can be accomplished many ways (e.g. MPLS tunnels, Multi-topology routing, physical separation via multiple networks, etc). Within these different service separation techniques, the operator may want to be able to additionally signal QoS classes. Therefore, it may be desirable to introduce a 2-level hierarch of service context identification. A mechanism to support such hierarchy is described below.

5.3 Solution Components

To solve the fate sharing issue of multiplexing all BGP AFI/SAFIs on a single session, "multisession BGP" (ietf-idr-multisession-bgp) was invented. In this form of BGP peering, the multiplexing of the peering is moved to the transport (TCP) and there are different peering sessions based on AFI/SAFI or arbitrary BGP attributes. Therefore a corrupt PDU in one service peering session will not cause other services to be torn down to recover from the corruption. No change to the BGP protocol peering state machinery is required to enable this feature. There is no requirement for multiple loopback addresses to be used. There is minimal configuration to enable the feature and it is easy to comprehend, manage and activate a new BGP peering session as it is the same as a single session.

The multiple sessions can terminate on different processes for fault isolation and also potentially terminate on different processors for performance isolation. Therefore each service can be prioritized and converged in an operator's choice of order. This is related to interdomain QOS as classes of routes can be divided by service class (gold, silver, bronze, etc) and fault isolation, performance tuning and prioritized can be applied.

As noted above, BGP sends withdraw messages for each prefix, per AFI/SAFI, and potentially per service topology and QOS class. This results in slow interdomain convergence as each prefix has to be withdrawn and readvertised. Today, this can take tens of minutes if multiple peers or sessions go down simultaneously. It would be preferable if BGP could announce multiple paths for a given prefix, thus avoiding the need to readvertise the new best path. A recent extension to BGP called "add path" (ietf-idr-add_patth) solves this exact problem. This extension is also applicable in IBGP with Route Reflectors, where the same problem is faced.

In addition to the ability to send the redundant path for a prefix in both External and Internal BGP, we need a faster protocol mechanism to announce failure conditions to trigger the use of the new path. An extension to BGP will be proposed (in the IDR working group of the IETF) called "Withdraw of Multiple Destinations". This extension will enable a single protocol message used to withdraw all prefixes from a specific peer, or to withdraw only those prefixes that match a specific pattern.

In sum, with these extensions we can now enable BGP to perform extremely fast reconvergence upon a failure and still maintain service level agreements. Convergence is now in the single seconds or less vs. potentially tens of minutes.

5.4 BGP Service Context Marking

We propose that a context capability should be used in combination with the multiprotocol capability to describe each destination (service) context. When two BGP speakers have exchanged their context descriptions (via opaque values), prefix exchange can happen using this special (service) context marking. The advantage of this approach is that the existing update message format can be reused, but still adding the benefit of advertising flexible descriptions of the destination tables and allowing updates targeted to these specific service forwarding tables. This can be done without changing the current update format in such a way in which all existing features that rely on the AF/SAFI pair to describe a forwarding table would be backwards compatible.

5.5 Context Exchange Procedure

When a BGP speaker wants to exchange routes using the new service context functionality, the speaker sends the context capability to its peer. The context capability itself lists each context it wants to use with a context identifier, length and description. Thus, a context for VOD (Video On Demand) service may be advertised as "42" with

complete independence of the actual packet markings. What is being exchanged is that the routes reachable for the VOD service are all marked with the opaque value "42." If there are multicast prefixes, VPNs, IPv4, IPv6, etc these additional services or reachability information can also be exchanged with the "42" context, without any change. The ID itself is opaque and does not define local or global QOS semantics. Instead it defines a service that is reachable and advertised by a peer. One could imagine that there would be, for example, a context value for the "low latency" service defined elsewhere in this document. That value could either be well known, or negotiated on a pairwise basis by two peering providers.

The Description Types may look something like this:

Description Types: 1: AFI (IANA AFI values) 2: SAFI (IANA SAFI values) 3: TOPOLOGY (0-255) 4: QoS (0-255)

Thus, an operator can now offer 256 QOS codepoints within up to 256 overlay topologies. This is considered to be beyond the current scaling needs but allows for future proofing and enables memory boundary alignment for the protocol attributes.

5.6 Summary

It is not considered to be necessary to signal anything beyond reachability and AS hop count. Again, BGP is not particularly good at passing dynamic data or link attribute information therefore, it is not recommended that we attempt to signal any of this information. History has also teaches us that global BGP route selection metrics are hard to agree on; hence, no change in selection metrics being advocated here. We are advancing that BGP is good at carrying around bags of data that the protocol doesn't care about. Our recommendation is that we use BGP to:

- 1) Exchange QOS and Topology information in an opaque manner to enable service differentiation
- 2) Extend the protocol that follows current BGP configuration, policies and management via a backwards compatible technique
- 3) Enable BGP with fast convergence features for per service "SLAs" a. Announce multiple paths per prefix/service b. Withdraw multiple prefixes per AFI/SAFI/Topology/QOS class in one message
- 4) Avoid interference with deployed features or availability mechanisms a. Remove fate sharing of services b. No changes to route refresh, graceful restart, etc.

6 Securing QoS

6.1 Motivation

In order to provide high quality service to specific customers, it is necessary to secure the network infrastructure as well as the use and provisioning of the service. What to secure

and how to secure it depend on what is done and how it is done (i.e., how the network is operated and what services are offered). For example, if all signaling and provisioning is done via manual configuration, then securing the network may be limited to securing the protocols used for configuration, as well as maintaining an audit trail of operator actions (e.g., to protect against insider attacks). Thus, this section is more a set of considerations to be taken into account.

6.2 Areas which need to be secure

There are multiple areas that need to be secured, including:

1. Securing the network infrastructure to ensure high availability of the network.
2. Securing the customer site
3. Securing the use of preferential services

The first two of these are critical to ensure that services are available and operate correctly, but are outside of the scope of this paper. Methods for securing the network infrastructure are for example being worked on in the IETF opsec working group (Operational Security Capabilities for IP Network Infrastructure, see http://www.ietf.org/html.charters/opsec-charter.html) and rpsec working group (Routing Protocol Security Requirements, see http://www.ietf.org/html.charters/rpsec-charter.html). Methods for securing a customer site are not currently the subject of standards efforts, but are the purpose of a variety of products such as firewalls and intrusion detection and/or prevention devices. A survey of current practices for securing service provider networks can be found in [OPsecPractices]. A survey of standards efforts related to network security can be found in [SecurityEfforts]. A set of best practices for cyber security and physical security can be found at www.nric.org, by clicking on "NRIC Best Practices", and then searching on the keyword "Cyber Security" or "Physical Security", respectively.

The set of practices and guidelines for network security, is constantly changing and evolving. Network operators must constantly be reviewing them and altering their procedures and practices accordingly.

Another general security issue is the design of protocols and the implementation of the protocols in software and hardware. This issue is also beyond the scope of this paper.

There are two broad areas of security that apply to IP-QOS: (i) Provisioning Security; and (ii) Service Security. Provisioning is the mechanism by which services are created and managed. Provisioning Security is how those mechanisms are protected against attack. A Service is some kind of TOS which is available to a subset of users (and their packets) in a network. Service Security protects that Service.

6.3 Provisioning Security

The goal of "Provisioning Security" is to secure the protocol aspects of the provisioning system, that is, the transfer of Provisioning Information between network elements. Provisioning Information includes, but is not limited to,

- QOS parameters such as bandwidth and latency, and
- traffic signatures, such as the DSCP

Routers, switches, network management stations, and end nodes all comprise network elements.

An ISP must also secure its network management elements and provisioning data (configuration files, audit trails, logs, and so on). If an NMS or configuration data are compromised, then the attacker can alter the TOS provisioning. If audit trails and logs are compromised, usage and billing data could be lost. Securing these elements is the same as general end-system and data-file security and, as such, is beyond the scope of this note.

There are also manual activities with regard to provisioning (business development people negotiating to create an IP-QOS, operators cooperating to implement and debug it, and so on). These activities can be vulnerable to attack and therefore must be secured, but discussion of these attacks and security mechanisms is beyond the scope of this paper.

Details of security (e.g. protocols and algorithms) are dependent on the exact protocols, algorithms, and procedures that provisioning uses. As such, these details are beyond the scope of this document. Instead, we concentrate on the requirements of security, talking about possible vulnerabilities, threats and attacks.

[Comment: Some OAM protocols don't currently have any authentication defined (e.g., LSP Ping). Others have defined authentication mechanisms, but these mechanisms are not yet available in products (e.g., BFD). Still others have security widely available (SSH, SNMPv3). I am not sure whether we should say more about the importance of authentication of OAM protocols.]

6.3.1 Goals

There are three goals of Provisioning Security:

1. Protection against unauthorized or inappropriate provisioning.

Attackers and other unauthorized parties must not be allowed to install services in a provider's network. They must also be prevented from altering, deleting, or otherwise reconfiguring existing services. A primary technique is to use cryptographically strong authentication.

2. Protection against DoS attack

Attackers and other unauthorized parties must be prevented from attacking the provisioning protocols in ways that prevent legitimate provisioning protocol operations from being performed.

3. Non-repudiation of provisioning requests

Insofar as provisioning represents a business relationship between two providers, with concomitant financial considerations, it is necessary that provisioning operations can not be repudiated. That is, if Bob sends a valid provisioning protocol operation to Alice, Bob must not be able to deny that he sent the operation.

6.3.2 Attacks

There are a number of attacks to which protocols in general are susceptible [RFC3552]:

- Eavesdropping
- Replay
- Message Insertion
- Deletion
- Modification
- Denial of Service

It is tempting to say that a particular attack is not of concern because the protocols in question will be used only in a way that obviates that attack, or the underlying network technology is such that the attack can not happen. We reject this reasoning. Protocol use and network topology have consistently evolved in ways that were quite unforeseen by the original designers.

The following subsections contain comments on each of the attacks.

6.3.2.1 Eavesdropping

Protection against eavesdropping is not necessary for safe operation of IP-QOS. It may be necessary or desired in order to prevent commercially sensitive information from being disclosed to a third party.

This non-requirement presumes that the provisioning protocols do not do things like carry cleartext passwords.

6.3.2.2 Replay

A replay attack is one where the attacker makes a copy of packets on the network and then retransmits them. Provisioning protocols must be safe from this attack.

6.3.2.3 Message Insertion

A Message Insertion attack is when an attacker creates a new message (or messages) and transmits it to the target. The provisioning system must protect against this as it could be used to send messages that alter or destroy existing services, or create new (unauthorized) ones.

6.3.2.4 Deletion

Message Deletion attacks are when the attacker prevents the proper reception of a message. Most good protocols are not very susceptible to this attack as the deleted message would appear as if the network lost the packet for other ("good") reasons. Well designed protocols will detect lost messages and retransmit them. If subsequent packets continue to be lost, then a failure of the communication channel will be detected and brought to the attention of network operators.

6.3.2.5 Modification

If an attacker can intercept, alter, and retransmit a message, then it is a modification attack. These attacks can be used to alter a provisioning request. Provisioning protocols should protect against this form of attack.

6.3.2.6 Denial of Service

By denial of service attack, we mean attacks against the provisioning system that prevent the provisioning system from working. These attacks can take a couple of forms

1. Flooding

Flooding DoS attacks work by simply sending so much traffic to the target that it spends so much time, memory, and so on, receiving, queuing, processing, and discarding the traffic that it has no resources left to process good traffic.

2. Algorithmic

These attacks utilize a weakness or vulnerability in the provisioning protocols (such as the TCP Timestamp vulnerability [CERT637934]).

A particularly insidious DoS attack can occur if the protocol uses cryptographic techniques to secure the packets. Cryptographic algorithms typically require significant amounts of resources. Thus, an attacker could overload a router's processor by sending a relatively moderate number of packets, each of which consumes a fairly large amount of resources to discard. The target could spend all of its time evaluating and discarding these packets. All other services provided by that target would then be effectively disabled. This attack can even occur indirectly. If some other protocol is attacked in this manner (e.g., BGP with MD5 authentication), there in some cases there might not be enough resources available to process provisioning protocol messages.

Some provisioning protocols make use of Soft State that needs to be periodically refreshed. If the refresh does not happen, the state is discarded (and thereby, the IP-QOS). An attacker can prevent that refresh. It could overload queues or the processor in the target. It could also prevent the refresh packets from reaching the target (e.g., by corrupting them in the network).

6.3.3 Security of Provider-Provisioned CE Devices

Where the service provider manages CE based devices, the service provider cannot ensure the physical security of the CE device. This leads to the possibility that a physical breach of security could occur at the customer site, leading to a possible mis-configuration of the CE device (for example, if a hacker where to obtain access to the console port of a CE router). The CE device therefore can not be trusted.

6.3.4 Carrier of Carriers Issues

In some cases a service provider may make use of services provided by a different service provider in order to interconnect their network. This is common in at least two situations: (i) where the carrier of carriers service is used to interconnect backbone routers in a service provider; (ii) Where the carrier of carrier service is used to interconnect a customer site with a service provider network. In this case the data plane and control plane may both be extended across the carrier of carrier's service.

In many cases, the carrier of carrier's service may be provided through use of virtual private network services (for example see [BGPMPLSVPNs]). Security issues with VPN approaches are discussed in the VPN Security Framework [RFC4111].

6.4 Service Security

"Service Security" means protecting the service itself from attack, abuse, and misuse. It is essential to protect the network from unauthorized use of premium services. For example, unauthorized use has the potential of defeating the provisioning efforts that are necessary for ensuring premium services.

[Comment: There may be some overlap between this section and the QoS section. Probably we need to first write the two sections, and then see whether any editing is appropriate to limit the overlap.]

In order for a packet to receive a particular service, that packet must be explicitly identified in some way. On ingress to the network, the packet may be identified by incoming interface (for example, some interfaces may for example receive best effort service for all traffic, while other interfaces may for example receive premium service for all traffic). Alternately the appropriate QoS for a packet may be identified by DSCP. Where MPLS is run across network boundaries the QoS for a packet may be identified by LSP, or by EXP.

Routers must be able to examine packets and determine whether they are requesting a particular service or not (and if so, which one) without significant performance degradation. If they can not do so, then the service is subject to attack by simply flooding a router with too much traffic for it to examine.

Providers must check the inbound traffic on all peer-provider-facing and customer-facing links. They must be able to detect and either discard, or remark as best-effort, traffic that enters from a peer-provider or customer that is not allowed to send such traffic.

A router must be able to properly internally partition traffic based on the service that the traffic should get. Best efforts traffic must not interfere with other traffic.

A Service may have a "maximum amount of traffic" associated with it. For example, on a specific interface some specific amount of traffic may qualify for premium service. The provider's routers must be able to enforce these limits. Traffic that exceeds the limits (but otherwise is eligible to receive the service) is called out of profile. There are a number of options for dealing with out of profile traffic:
- discard it
- mark it as best effort
- mark it as out of profile and requesting the service. This traffic could be treated as best-effort, or given the service. (mark is used 'generically' -- it could mean changing the existing header or (re)encapsulating the packet).

The policing tests must be low-cost. If policing is too expensive (i.e. causes significant performance degradation) then it is possible to attack the policer by flooding it with packets.

A service provider can not trust that a peer service provider has adequate security. Thus, service security measures must be provided on inter-provider links.

6.5 Security Guidelines

This section is a brief list of procedures and practices that network operators should follow.

1. Be in contact with, understand, and constantly review all available security practices, guidelines, alerts and other pertinent information. The nature of security threats and the methods for dealing with them is constantly changing. Network operators must constantly adapt their own security procedures.

   Good sources of security information include CERT, NRIC, the IETF and NANOG.

   Operators must also review all security-related announcements and information available from their equipment vendors. Security patches should be installed as soon as practical.

2. Do not rely on cleartext passwords and the like. Assume that all network traffic is subject to sniffing and analysis. Cryptographically strong algorithms must be enabled and used. This is critical for network management protocols and service provisioning protocols.

   Whenever packets/messages/operations fail the failures must be counted and logged. Security personnel should be notified and take appropriate actions. One should never ignore a small violation as "one of those things". Large attacks start as small probes.

3. Do not trust customer networks.
   You can not assume that the customer's security practices are good. The customer could easily generate excessive traffic for a particular service. Even if the customer's CE device is provisioned and/or managed by the provider. Since the device is not under the physical control of the provider, it can be it can be reconfigured or otherwise compromised.

4. Do not trust peer networks.
   Just as a customer's net can be compromised, so too a peer provider's network can be compromised. Security practices which are deployed on links facing customers must also be deployed on links facing other providers.

5. Filter & drop traffic that comes from a place where it shouldn't. If a peer or customer is not supposed to be sending you traffic for a particular service, do not accept packets from that peer or customer that request the service. This might just be a routing or configuration issue on the part of the peer or customer, but it could also be an attack.

   This is especially critical for management and provisioning protocol traffic.

6. Filter and Rate-limit ingress traffic
   The best mechanism to ensure that a service is not attacked is to detect all packets that are to get that service and rate-limit them at the point they enter the network. Packets which are in violation of this limit may either be dropped or remarked as nonconforming or "not to receive the service". Which mechanism to use depends on the business agreements and the service being requested.

   Selecting the rate at which the traffic is limited is complex. Factors include contractual obligations and available network resources. From a security perspective, we will assume that the network resources are available to meet the contractual obligations. Therefore, the rate limit should be no higher than the contractual obligation. This prevents someone from using "more than they should".

   Traffic that is not to receive the service also should be rate-limited. If the non-QOS traffic is "too much", it could constitute a denial of service attack.

7. Read, understand, and apply the practices in [OPsecPractices]. If you do not apply one of these practices, you should understand the practice, understand the vulnerabilities (if any) that you will create by not applying the practice, and have a good reason for doing so.

   Keep up to date with this document as it is revised.

8. Read, understand, and apply the practices in [SecurityEfforts].

   Keep up to date with this document as it is revised.

9. Read, understand, and apply the practices in [RFC3871]. This document spells out a number of practices and requirements for operators and network equipment. You should understand the extent to which any device you have deployed either meets the requirements or why it does not (understanding that there is no perfect device and that tradeoffs are needed).

   Keep up to date with this document as it is revised.

7 Operational Issues

- In general, PPI links are "aggregated interfaces," which contains a mixture of traffic from different customers.
- In other words, treatments are applied to packets in an aggregate fashion; no per-customer treatment is applied to packets, except under exceptional conditions, (e.g.: security-related issues).
- RFC2547, Option A, is the exception to this rule.
- While possible to use Layer-2 (e.g.: ATM PVC's, F/R DLCI's, etc.) to provider per-customer QoS guarantees, that is deemed far too onerous to set-up and maintain, particularly in light of, for example, 1000's of customers on 100's of peering circuits going between 10's of carriers.
- Discussion of whether NetFlow (cflowd) or interface counters are better for monitoring queue utilization, (relative to scheduler %'s and relative to accounting for billing purposes)? For instance, NetFlow may not be acceptable for billing data; but, is acceptable for monitoring for capacity planning or troubleshooting?
- Recommended intervals to monitor scheduler %'s vs. actual utilization?
- Need to ensure adequate capacity is always available to soak up demand from 1 or more other PPI failures. Likely need to write SLA's for timely delivery of additional capacity into 'peering agreements' to ensure both parties meet their obligations, for both PE-CE and ASBR-ASBR links.
- Shaping isn't used for Voice because it adds to jitter.
- Ability to use IP traceroute or LSP PING for debugging of drops, (mis-)marking, etc.?

At this stage, this section attempts to identify issues for a white paper guideline. Answers to the questions suggesting operational procedures and new questions are solicited.

The advent of interconnections where we undertake to deliver traffic with a specified quality under the condition that it conforms to certain conditions brings new operational challenges. These are related to the operation of the differentiated services enabled interconnections as such or to QOS related capabilities such as timely re-routing of traffic across domain borders or to functions supporting the business relationship of the interconnecting parties such as accounting functions.

This section is structured according to the FCAPS model. Some of the FCAPS topics central to interprovider QOS have been covered already in other chapters:

- Performance monitoring has been given extensive coverage in the measurement chapter.
- Policing, scheduling and dimensioning have been covered in the service class definition chapter.
- Fast rerouting is covered as part of the routing chapter.
- Security issues are covered in the chapter Securing QOS.

7.1 Fault

Fault management is not specific to interprovider QoS but the requirements on timely fault detection and service restoration are more stringent as a consequence of the availability guarantees. This means that fault detection and notification mechanisms and performance used between interconnected parties both in the control plane level and network management level must be agreed on as part of the SLA. This is valid both for the PPI interface and the CPI interface.

Fault isolation and troubleshooting may require a coordinated effort by the providers involved. To make the process efficient some prior agreement on the responsibilities of the providers regarding notification, troubleshooting and sharing trouble shooting information should be made.

The basic assumption is that each provider is responsible for troubleshooting their own domain. Therefore it should not be a requirement for a provider to react to active probes (traceroute and ping) other than on the PE and ASBR nodes.

In case of lost connectivity service availability will depend on the efficiency of rerouting traffic. Assuming the sender pays model each provider is responsible for rerouting the traffic within their domain and slow convergence will impact the availability metric in the SLA. This means that there is a direct connection between the requirement on fast rerouting of traffic and the formulation of the availability guarantee and that there is no need for any exchange of information on internal routing protocol rerouting performance.

In case of service affecting faults it is considered good practice to notify customers of expected duration of problems. This should be done via same channels as notification of service windows.

[Are temporary performance measurements, in addition to the regular ones, necessary during troubleshooting? Shall measurements be allowed/required between measurement points in the two different provider domains? How do we know that performance measurements are valid for the paths that our traffic is using?]

7.2 Configuration

Due to the higher demands on availability (or specified availability) there will be a need for correlating configuration events that might affect service performance.

Regarding configuration work on interprovider interfaces (PPI and sometimes CPI) there must be a common change process that minimizes the affect on customer traffic due to bad correlation. This process includes approval, planning and scheduling the work to be done while still allowing for urgent corrective action to be performed.

To allow for service affecting management activities to be performed on networks with a minimum of customer impact it is customary to define service windows when degradation or loss of service is accepted as being within the limits of the SLA. Due to the global scope and the number of different administrations may be involved in the interprovider QOS case it is not possible to schedule a regular service window that is suitable for everyone. As a consequence of this a provider that wishes to utilize a service window must notify all partners and customers ahead to give forewarning and to make sure that the intent of the definition of a service window is not abused.

Other providers may wish to take action as a consequence of the activation of a service window. This could be to notify their customers, rescheduling of some activities or to take precautionary action. To allow for efficient processes to be implemented the length of the notification period and other constraints such as the frequency and length of the service windows allowed need to be generally agreed upon between providers.

If a provider needs to perform urgent service affecting management it is considered best practice to give notification as early as possible even though this does not validate a service window.

Today notification is given via e-mail to an agreed contact point. In future we might wish to find other means of notification to cater for scalability and accountability.

[More details on service windows in the SLA section.]

7.3 Accounting

Although there may be different models to settle payment between providers exchanging QOS traffic it is a reasonable basic assumption that it is the receiver of the traffic, promising to deliver it with a defined quality of service, that performs a service to the sender. In other words it is the sender who pays the receiver.

Based on this it is the receiver of the traffic that will be responsible for measuring traffic volume per service class for billing purposes (in those cases where the actual traffic volume affects the billing). In order for the sending party to verify the measurements (if needed) this should be done using a well known and well specified method e.g. standard interface counters that may be applied both on the outgoing interface and the incoming interface on the PPI link.

7.4 Performance

On the PPI links there will be a need to agree upon how utilization is to be measured and the upgrading rules and process to use. In some cases there will be a clear customer provider relationship where the customer will have the responsibility to upgrade. In other cases (when there is a peering relationship) the need for upgrading might not coincide completely and must therefore be regulated. In the service definition chapter there is a proposal for how policing set back to back on the PPI link might be used for utilization monitoring.

It will be common practice to set up a number of interconnection points between two providers. These will be used as back up paths for each other. A provider might also wish to utilize several downstream providers in order to ensure high availability. A provider might choose to try to spread the utilization over the different paths or may prefer a certain path due to e.g. delay or cost reasons. This means that the network split of the load in case of failures is cannot be assumed to be known. To ensure that dimensioning of the networks (both interprovider links and the networks in general) is based on the correct information the back-up requirements (and possibly rerouting policy?) should be agreed upon between interfacing providers.

8 Other issues and Items Not Discussed

- This section of draft will be to identify topics that were considered but where no best practice recommendation was developed or were deemed beyond the scope of the white paper. Topics that may be included in Section 8 as "beyond scope":
  - Signaling issues. (Is OEM signaling part of this?) Including service discovery, how to identify new devices or locate users/new devices.
  - End-user equipment (DSLAMs/access routers)

8.1 *Maintenance Windows*

Providers of the real-time network class are expected to need similar maintenance periods as other providers. That is, every provider will have both planned and unplanned maintenance periods. Since industry practice does not consider planned maintenance outage as unavailability, planned maintenance periods should be considered separately. Unplanned maintenance should be considered as a component of unavailability.

In the case of a single provider, network performance objectives need not be met during planned maintenance. The service contract should make the hours clear, and whether

notification of a customer affecting activity is required, how much notice, etc. Providers may try to plan maintenance for low usage periods, say 2am-4am local time.

Extending SLAs to multiple providers is more complex. How can a customer-facing provider inform a customer of maintenance periods for traffic having a multitude of destinations which wind their way through multiple providers - each of which have their own planned maintenance periods? For global traffic, what is the likelihood of traffic crossing a provider who means well by doing planned maintenance during the "graveyard shift" when that traffic impacted may be for a customer's "busy hour"? It would be beneficial if planned maintenance notification could be extended to network partners, as well as customers, but how much value real or perceived is there for future or long lived sessions to have this foresight?

Inter-provider maintenance windows could be defined per path as the super set of all individual windows, providing that the result is acceptable to the customer. How windows match could be a key criterion to decide over which providers a path is routed. If end-to-end maintenance though a particular set of providers is unacceptable, an alternate set might be found.

A non-signaled static approach could only be statistical, possibly based upon heuristics, though this seems unlikely to satisfy customers.

Global agreement concerning a specific absolute time for when planned maintenance occurs is clearly impractical. However, there may be practical methods to coordinate within constraints. A notification scheme that is communicated to all potential affected parties seems to be the most practical and satisfactory. Providing the notification period and procedure was complied with, the planned maintenance could proceed. Any provider that has customers that were likely to be unreasonably impacted by another providers planned outage would have the right to negotiate changes to the requested window. In this case any changes agreed must still be communicated in accordance with the notification period and procedures to all other affected providers. This regime would require all notification requests to be cascaded through providers as one provider may not know what it used beyond the adjacent provider's network

Current industry best practice is for the communication of “planned maintenance” via electronic text (Email). This poses some cultural and language difficulties if the communication is written in the native language of the recipient. The format of the notice and its contents needs to be well defined to avoid any misunderstanding. No current industry standards have been identified for this. Planned maintenance periods could be signaled during session setup, during sessions, and/or indicated along with measurement exchanges, via a database using a standardized message structure.

A minimum notice of 15 days is recommended unless less it is otherwise agreed upon by all affected parties. For urgent work it is good practice, and the practice is encouraged, to give as much advanced notice as practically possible that a service impact is about to occur. Where outage notification is less than the recommended 15 days, then it is at the

discretion of the affected parties as to whether the outage is accounted as unavailability or "planned maintenance" for SLA reporting purposes. Provided the notice period (15 days) is adhered to, the notification would be accepted by all parties unless there were exceptional circumstances.

During both planned maintenance periods and periods of unavailability, the predicted resumption of service should be indicated to partners using the same communication channels.

We have not yet collected and analyzed sufficient issues, practices and potential solutions to this maintenance window aspect of Inter-provider QoS. Therefore we have no complete "best practice" proposal yet. This is an area for further study.

9 References

[AGG] Chan, K., Babiarz, J. and F. Baker, "Aggregation of DiffServ Service Classes", Work in Progress, draft-chan-tsvwg-diffserv-class-aggr-01.txt, February 2005.

[BGPMPLSVPNs] E.Rosen, Y.Rehkter, "BGP/MPLS IP VPNs", draft-ietf-l3vpn-rfc2547bis-03.txt, October 2004.

[Blake98] Blake, S., D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. *An Architecture for Differentiated Service.* RFC 2475, December 1998.

[CERT637934] CERT Vulnerability Note VU#637934, "TCP does not adequately validate segments before updating timestamp value", http://www.kb.cert.org/vuls/id/637934

[CIPPM] Mahdavi, J. and V. Paxson, "IPPM Metrics for Measuring Connectivity", RFC 2678, September 1999.

[CLASSES] Chan, K., Babiarz, J. and F. Baker, "Configuration Guidelines for DiffServ Service Classes", Work in Progress, draft-ietf-tsvwg-diffserv-service-classes-00.txt, February 2005.
[DVIPPM] Demichellis, C. Chimento, P. IP Packet Delay Variation Metric for IP Performance Metrics (IPPM). RFC 3393, November 2002

[FRIPPM] Paxson, V., Almes, G., Mahdavi, J. and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, May 1998.

[IETF RFCs that provide basis for the definition of metrics need to be added]

[OPsecPractices] Merike Kaeo,"Operational Security Current Practices," draft-ietf-opsec-current-practices-01, July 2005.

[OWIPPM] Almes, G., Kalidindi, S. and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[PLIPPM] G. Almes,S. Kalidindi,M. Zekauskas A One-way Packet Loss Metric for IPPM. RFC 2680, September 1999.

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option," August 1998

[RFC3552] Rescorla, E., and B. Korver, "Guidelines for Writing RFC Text on Security Considerations," July 2003.

[RFC4111] Luyuan Fang, "Security Framework for Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC4111, July 2005.

[RosRek05] Rosen, E. and Y. Rekhter. *BGP/MPLS VPNs.* Work in Progress, draft-ietf-l3vpn-rfc2547bis-03.txt

[RosRek99] Rosen, E. and Y. Rekhter. *BGP/MPLS VPNs.* RFC 2547, March 1999.

[RTDIPPM] Almes, G., Kalidindi, S. and M. Zekauskas, "A Round-trip Delay Metric for IPPM." RFC 2681, September 1999.

[SecurityEfforts] C. Lonvick and D. Spak, "Security Best Practices Efforts and Documents," draft-ietf-opsec-efforts-01.txt, July 2005.

[Y.1541]