# Inter-provider Quality of Service

White paper draft 1.1

November 17, 2006

a white paper prepared by the

Quality of Service Working Group

MIT Communications Futures Program (CFP)

Participating Contributors included


Shane Amante (Level3)
Nabil Bitar (Verizon)
Nils Bjorkman (Teliasonera)
Ross Callon (Juniper)
Kwok Ho Chan (Nortel)
Dave Clark (MIT)
Anna Charny (Cisco)
Bruce Davie (Cisco)
Dave McDysan (Verizon)
Luyuan Fang (Cisco)
Phil Jacobs (Cisco)
Frank Kastenholz (Independent)
Roman Krzanowski (Verizon)
William Lehr (MIT)
Xiao-Gao Liu (Nortel)
Kevin Mason (TNZ)
Jaime Miles (Level3)
Henrik Villfor (Teliasonera)
David Ward (Cisco)


The opinions expressed in this paper are drawn from consensus views among the working group's participants, and do not represent official views or policies of CFP's sponsoring companies or universities.

The ***Communications Futures Program*** (CFP) is a partnership between university and industry at the forefront of defining the roadmap for communications and its impact on adjacent industries. CFP's mission is to help our industry partners recognize the opportunities and threats from these changes by understanding the drivers and pace of change, building technologies that create discontinuous innovation and building the enablers for such innovation to be meaningful to our partners. Further information about CFP, see http://cfp.mit.edu.

**Table of Contents**

# Executive Summary

This document presents a 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 [RFC 2475] is now widely deployed within the networks of single providers. This is especially the case for providers of network-based VPNs (see, for example [RFC 2547, RFC 4364]). Some providers are now beginning to interconnect with each other via "QoS-enabled peering" 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, delay variation and loss are defined consistently.

**Service Class Definition.** The "low latency" 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). Although this document only outlines detailed criteria for a single classes of service beyond best effort, its goal is to remain flexible so that additional classes of service may be added. Furthermore, any individual Service Provider is free to offer additional service classes beyond those defined here.

**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.

# 1 Introduction

Quality of Service (QoS) technologies based on the Differentiated Services architecture [RFC 2475] are now widely deployed within the networks of many service providers. This is especially the case for providers of network-based VPNs (see, for example [RFC 2547, RFC 4364]). Some providers are now beginning to interconnect with each other via "QoS-enabled peering" 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 have proven 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.

## *1.1 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 better 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 RFC 4364 (BGP/MPLS VPNs, formerly known as 2547 VPNs[1]). This document will

[1] RFC 2547, which was the informational RFC that described MPLS/BGP VPNs, has now been superseded by the standards track RFC 4364.

not restrict itself to BGP/MPLS VPNs - any IP VPN service should be supported - but we will address the specific QoS issues of interconnecting providers of BGP/MPLS VPNs, including the MPLS-based interconnection styles (referred to as options (B) and (C) in [RFC 4364]).

### *1.2 Relationship to standards*

This work draws heavily on the efforts of both the IETF (particularly the IPPM working group) and the ITU (particularly the Y.1541 recommendation on service classes). Where possible we have tried to be consistent with these efforts, but there are a few points on which we have diverged. In some cases differences arose from the authors' desire to make recommendations that could be implemented with existing equipment at acceptable cost – notably in the case of probing frequency for measurements. Our focus on practical methods of concatenating services across multiple providers' networks led to a particular definition of delay variation that differs from Y.1541.

It is possible that some parts of this white paper will be used as the basis for standards contributions in the future. In that case it will no doubt be necessary to revisit the differences between this paper and the current standards and to consider possible compromises.

## 2 Reference model & terminology

### *2.1 Definitions*

**Access :** That part of an end to end connection from the customer's side of the CE router to the customer's side of the first PE router.

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

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

**Core**: That part of a provider's network from the customer side of the PE router to the customer's side of the distant PE router, or the mid point of the ASBR – ASBR provider to provider interface.

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

**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, with each site connecting to one or more VPN Service

Providers. In the context of the Internet, a customer typically connects to an Internet Service Provider at one or more locations.

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

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

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

**Option A (B, C):** Methods for interconnection of MPLS VPNs across service provider (and AS) boundaries, defined in RFC 4364

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

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

**PPI: Provider to Provider Interface.** The interface defined by a single, physical link between two, different 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.

**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.

**Unmanaged CE**: A Customer Edge router that is managed by the customer, rather than by a provider.

Acronyms used herein include the following:

| | |
|---|---|
| **AFI** | Address Family Identifier |
| **AS** | Autonomous System |
| **ASBR** | Autonomous System Border Router |
| **ATM PVC** | Asynchronous Transfer Mode Permanent Virtual Circuit |
| **BGP** | Border Gateway Protocol |

**CDF** Cumulative Distribution Function
**CPI** Customer to Provider Interface
**DSCP** DiffServ Code Points
**E2E** End to End
**EF** Expedited Forwarding
**EXP** Experimental (field in MPLS header that carries Class of Service information)
**FCAPS** fault-management, configuration, accounting, performance, and security
**FR DLCI** Frame Relay Data Link Connection Identifier
**GigE** Gigabit Ethernet
**IP** Internet Protocol
**IPPM** IP Performance Metrics
**IP-QoS** IP Quality of Service
**IPVPN** Internet Protocol Virtual Private Network
**ISP** Internet Service Provider
**LL** Low Latency
**LSP** Label-Switched Path
**MPLS** Multi-Protocol Label Switching
**NMS** Network Management System
**NLRI** Network Layer Reachability Information
**OAM** Operational and Management
**OPSEC** Operational Security Capabilities
OWD One Way Delay
**OWJ** One Way Jitter
**PDU** Protocol Data Unit
**PHB** Per Hop Behavior
**PPI** Provider to Provider Interface
**PWE3** Pseudo-wire emulation edge to edge
**RFC** Request For Comment
**RPSEC** Routing Protocol Security Requirements
**SAFI** Subsequence Address Family Identifier
**SLA** Service Level Agreement
**SP** Service Provider
**TOS** Type Of Service
**TWAMP** Two-way Active Measurement Protocol
**VC** Virtual Circuit
**VoIP** Voice over IP

## *2.2 Reference approach*

The key underpinning recommendation of this paper is that the DSCP value in the IP header is the default definitive indicator to all providers in the end to end path of the QoS treatment an IP packet should receive. Providers may use parameters in other protocol headers to convey QoS treatment required (e.g. where encapsulation of the IP packet

occurs) but in the event that these marking differs from the QoS class indicated by the DSCP parameter, the later will be the definitive indication.

## *2.3 Reference model*

For simplicity, we consider first the single provider case depicted in Figure 1. In this model, as in [RFC4364], 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. We define the Customer to Provider Interface (CPI) as the link between the CE and the PE. In general, the CPI is arranged so that only one customer sends traffic on a given CPI (this may involve some multiplexing layer such as Frame Relay DLCIs.) The CPI represents the typical boundary of trust between the provider and the customer. That is, the provider does not trust the customer to mark packets correctly or to send at a certain rate – this fact influences policing, for example, as discussed below. 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 provider/customer contract known as a Service Level Agreement (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.

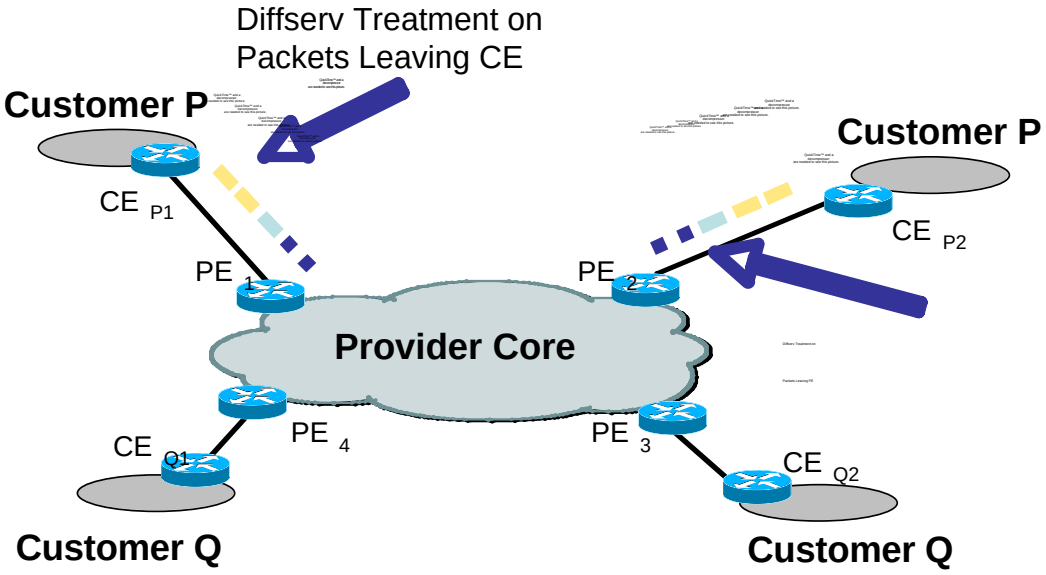



**Figure 1. Basic Diffserv Model for Single Provider**

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. We define the link between the two providers as the Provider to Provider Interface (PPI).

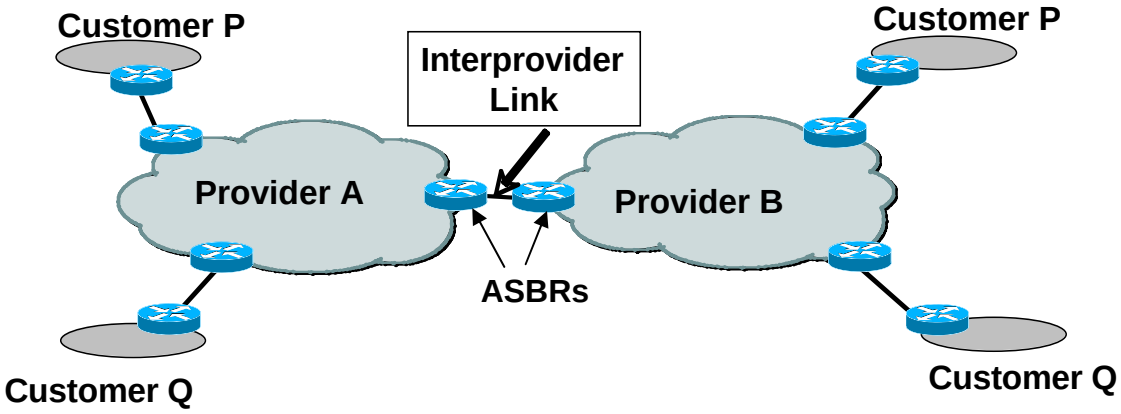



**Figure 2. Simple Interprovider Topology**

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, and the providers may have different markings for a given 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. QoS 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 RFC4364 VPNs) as well as MPLS interconnects of both option B and option C styles [RFC4364]. Interconnection using MPLS traffic engineered LSPs should also be possible. It should also be possible to support any sort of layer 2 interconnect (e.g. ATM, Ethernet, etc.). 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. For example, a customer may have an SLA that allows him to send traffic with DSCP=X up to a rate *r*, with excess packets being downgraded to best effort. However even if the packets are treated as best effort by the provider, the customer wishes to retain the DSCP marking of X 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, while leaving the customer's headers intact.

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 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. In option A or pure IP interconnection it may be possible to encode the marking in a layer-2 dependent way to again leave the customer header unchanged. 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, QoS 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 service class, targeted for the transport of Voice over IP (VoIP) and other latency-sensitive applications. We call this service the "Low Latency" class. It is assumed that this service is offered in addition to the standard "Best Effort" class. We focus on a single additional

class not because we think that two classes are always sufficient, but because most of the issues that need to be tackled become apparent as soon as one tries to go beyond a single best effort class for all traffic.

Note that a service class is defined in terms of "black-box" behavior – that is, we define the externally visible attributes of the service (e.g. loss, delay) rather than internal implementation mechanisms (e.g. Diffserv PHBs). In this respect service classes are similar to the Diffserv concept of a Per-Domain Behavior (PDB) [RFC 3086], but we do not limit the scope of a service to a single domain. To be precise, a service class is defined by the externally visible treatment that the packets in that class receive as they traverse a network (in terms of loss, delay, and delay variation, and potentially policing). There may be additional aspects of a service class definition such as a default marking for packets in that service class.

### 3.1.1 "Low Latency" (LL) Service Class

The additional service class defined in this document is the "Low Latency" (LL) Service Class, which is to be used for the transmission of services that require low delay, low delay variation and low loss between two, or more, disparate carriers. The class is intended to be suitable for real-time applications such as VoIP, but there is no restriction on which applications may actually use the service. Mapping of applications to service classes is left to customers.

For many applications, the LL service class must carry traffic bi-directionally (e.g. the associated signaling and media packets for both directions of an individual VoIP call). However, there is no requirement to provide a symmetric path for the bidirectional traffic flow between any given source and destination.

### 3.1.2 "Best Effort" (BE) Service Class

The Best Effort Service class is the default service class that is assumed to be available everywhere. Traffic that has not been explicitly identified and associated with another service class will receive Best Effort treatment. The Best Effort service class typically has no guarantee with respect to latency, delay variation, or packet loss; however, carriers typically do endeavor to provide for satisfactory delivery of packets in this service class, and SLAs for best effort are not uncommon.

### 3.1.3 Other Service Classes

Service providers are at liberty to offer any number of service classes above and beyond those defined in this document. Indeed it is typical to offer four or more service classes to end users and also to use one or more internal classes for network control (e.g. routing protocol) traffic. We expect that more classes will probably need to be agreed upon for interprovider use at some point in the future, but we have deferred the discussion of additional classes for now. As noted above, we believe that even agreeing on how to

support one additional class (i.e. the Low Latency class) beyond the standard best effort class raises many if not most of the hard problems that need to be addressed.

We also note that, even if there were a larger number of "standard" service classes that could be offered in an interprovider context, it is likely that providers would continue to offer some additional classes beyond the standard set as a means of competitive differentiation. An interprovider QoS model should allow for such flexibility.

## *3.2 Customer to Provider Interface (CPI) Behavior*

### 3.2.1 Marking of Customer Traffic

At the CPI, the customer must appropriately mark packets that are to receive Low Latency service. This document proposes that the default DSCP for EF (101110) should be used for packets that the customer intends to receive Low Latency service [RFC3246]. If packets at the CPI are MPLS encapsulated (e.g. because a Carrier's Carrier service is being offered to the customer) then the top MPLS header should contain an EXP value of 5.

For traffic that is to receive best effort service, the customer should mark the packets with a DSCP value (or EXP value) of 0.

Providers are free to specify the use of other DSCP or MPLS EXP values for other service classes beyond Best Effort and Low Latency.

There is no restriction as to what type of traffic the customer may place in any service class. For example, if the customer chooses to place bulk data traffic with long packets in the Low Latency service, it may degrade the performance of that customer’s voice traffic experiences, but that is up to that customer to decide.

If traffic from the customer is marked with a DSCP or EXP value that does not match any of the acceptable values that have been agreed upon as part of the customer's SLA, the provider may take any action that the provider considers appropriate (such as dropping or remarking). Note that this issue also arises at the PPI and is discussed below.

### 3.2.2 Policing and Re-Marking

Policing of the Low Latency class is performed at the CPI as described in Figure 4 of [CLASSES]. That is, the SLA includes a token bucket rate and burst size; traffic sent by the customer that exceeds this token bucket at the CPI will be dropped. Such policing must be performed at the PE in the case of unmanaged CEs. It may be performed at the CE if and only if the CE is managed by the provider.

Re-marking of excess traffic may be appropriate for future service classes, but is not recommended for the Low Latency class (see section 2.8 of [RFC3246]).

The configuration of egress queuing (from egress SP's PE to ingress CE) is a local matter for the SP. It is also a local matter for the provider to decide if he wishes to use one of the pipe models of [RFC3270].

## *3.3 PPI Behavior*

### 3.3.1 Marking of Traffic at PPI

At the PPI, packets that are to receive Low Latency service must be appropriately marked. This document proposes that the default DSCP for EF (101110) should be used for packets that are to receive Low Latency service [RFC3246]. If packets at the PPI are MPLS encapsulated (e.g. because options B or C are in use at the PPI) then the top MPLS header should contain an EXP value of 5.

For traffic that is to receive best effort service, packets should be marked with a DSCP value (or EXP value) of 0.

Providers are free to negotiate with their peers the use of other DSCP or MPLS EXP values for other service classes beyond Best Effort and Low Latency.

If traffic from one provider to another does not match one of the agreed-upon DSCP or EXP values for that interface, then the behavior is unspecified – that is, traffic may be dropped, remarked, or transmitted unmodified with any QoS the receiving provider chooses.

### 3.3.2 Policing and Re-Marking at the PPI

Policing of the Low Latency class is performed at the PPI as described in Figure 4 of [CLASSES]. That is, the SLA between the peering providers includes a token bucket rate and burst size; traffic sent by a provider that exceeds this token bucket at the PPI will be dropped. Such policing must be performed at the ASBR of the receiving provider. It is expected that the token bucket parameters will be statically configured as a result of offline configuration.

Re-marking of excess traffic may be appropriate for future service classes, but is not recommended for the Low Latency class (see section 2.8 of [RFC3246]).

On receipt of packets from the PPI, an SP may encapsulate packets, using either IP or MPLS, and mark the encapsulating header with a 'local-use' DSCP or EXP values within that provider's backbone, as long as the encapsulated header is not modified.

It is generally considered desirable to avoid remarking of customer's traffic in a way that the customer can detect, i.e. by modifying the customer's DSCP values. This means that

if remarking is required for some reason (e.g. to deal with unknown or unexpected DSCP values) it is desirable to encapsulate the customers traffic with a header that can carry the desired marking, rather than modifying the customer's DSCP. The implication of this policy at the PPI is that it is preferable to carry traffic in an encapsulation that supports some sort of marking other than the customer's DSCP. Option B and Option C meet this requirement, since there is an MPLS header to carry the marking. It is also possible to apply an MPLS (or IP) header at the PPI purely for the purposes of carrying an EXP (or DSCP) value – this is feasible even with Option A or a pure IP interconnect.

The configuration of egress queuing (from one provider's ASBR as he transmits onto the PPI link) is a local matter for the SP. It is also a local matter for the provider to decide if he wishes to use one of the pipe models of [RFC3270].

## *3.4 Definitions of metrics*

The measurement method proposed is active probing, which is the generation and measurement of synthetic traffic designed to model the performance of aggregate customer traffic. The metrics discussed in this section relate both to customer traffic and active probes.

### 3.4.1 Initial Considerations

The Low Latency service class is characterized by three network performance metrics: one-way latency, one-way packet loss, and one-way delay variation. In general, the metrics follow the approaches defined in the IPPM group at the IETF. The main challenge in the context of this work is to restrict the options available, as the IPPM RFCs allow a great deal of latitude. Since our desire in this work is to produce service classes with metrics that can be meaningfully concatenated, it is important to have reasonable commonality of metrics across providers.

As much as possible, we have tried to be consistent also with the definitions of metrics in Y.1541. However, in some cases where the practical feasibility of following the definitions precisely was in question, we have recommended some slight variations (e.g. in the choice of measurement frequency). We are also particularly concerned with the problem of meaningfully concatenating the metrics across multiple providers, and this has motivated a slightly different definition than that chosen by the ITU in some cases.

Additional metrics can also be defined for the low latency traffic class but their use is not required by this document. These include: availability, connectivity, throughput, and packet reordering.

There is a widespread practice of reporting two-way metrics or one-way metrics derived from two-way measurements. However, our preference is for one-way metrics, as they reflect most accurately the performance of the network. One-way measurements do, however, require accurately synchronized clocks. This document proposes that one-way metrics should be reported whenever possible; one-way values derived from two-way

measurements may be used only if one-way measurement is impossible, and the fact that they are not true one-way metrics must be reported.

All performance guarantees are only for conforming packets/traffic – packets sent outside the SLA (token bucket) parameters for a given interface are not counted in SLA measurements of their senders' service class, even if they are delivered.

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, operational performance measurement systems. Single measurements are used and reported during the debugging or calibration process.

For the Low Latency service class, all metrics should be defined for packets that are representative of the traffic that will use that class. Thus they should use IP/UDP/RTP packets with a payload size of 160 bytes (representative of common VOIP codecs today). Test packets must also be marked with the appropriate DSCP or MPLS EXP values as defined above for the Low Latency class.

For all the metrics defined here, a number of measurements must be taken over a defined time interval. When reporting these measurements, the time of the start of the interval should be reported, relative to UTC.

For all the metrics defined here, we have recommended a sampling frequency of 200ms and a measurement interval of 5 minutes, leading to 1500 samples per interval. Choosing the sampling frequency is clearly a complex tradeoff between accuracy and load on the network itself and on the measurement devices. The authors believe a 200ms sampling interval is a reasonable compromise, and we note that providers may probe more often if they wish (perhaps on an exceptional basis, e.g. for troubleshooting.) See also Section 4.4.1.1 for discussion of the precision of delay measurements in particular.

Test packets for all the metrics defined here should be generated by a Poisson process to avoid any periodic effects. Providers may choose to use uniform test packet intervals but this must be clearly stated with any metrics reported.

### 3.4.2 One-way Delay

The definition of one-way delay (OWD) follows the approach defined in RFC 2679.

The single instance of the one-way delay measurement is defined as the time the test 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.

The OWD metric is reported as the arithmetic mean of several (specified) single measurements over a specified period of time. Errored packets and lost packets are excluded from the calculation. The metric is reported to 1ms accuracy, rounded up, with a minimum value of 1 ms.

- The recommended maximum evaluation interval = 5 minutes
- Recommended mean packet separation=200ms. (Providers are at liberty to measure more often than every 200ms and to report that fact)

### 3.4.3 One-way IP Packet Delay Variation [IPDV]

A number of definitions of IP Packet Delay Variation (IPDV) have been proposed debated. (See Appendix C for further discussion of different approaches). This document proposes the following definition for IPDV, following the approach defined in RFC 3393.

A definition of the IP Packet Delay Variation (IPDV) can be given for packets inside a stream of packets.

The singleton measure of IPDV for a pair of packets within a stream of packets is defined for a selected pair of packets in the stream going from measurement point MP1 to measurement point MP2.

In this document, the $IPDV_{(n)}$ is the difference between the one-way-delay of the selected packet and the packet with the lowest OWD in the evaluation interval.

$IPDV_{(n)} = OWD_{(n)} - OWD_{(0)}$

When reporting IPDV it is more practical and useful to report at least one point on the IPDV distribution in an evaluation interval rather than the entire distribution of singleton measures.

This paper recommends that the selected point of the distribution follow the ITU-T Y.1540/1541 IPDV definition according to section "6.2.3.2 Quantile-based limits on IP packet delay variation". Specifically, we recommend that at least the $99^{th}$ percentile of the IPDV distribution over a 5 minute measurement interval is reported. Furthermore, in the rest of this paper (with the exception of Appendix C) and unless noted otherwise, we will use the term “IPDV” indistinguishably with its 99-th percentile IPDV, and will frequently omit the subscript in $IPDV_{(99th\ P)}$ where it causes no confusion.

Appendix C presents a method where two points of the IPDV distribution are used instead of the 99-th percentile only.

- The recommended maximum evaluation interval = 5 minutes
- Recommended mean packet separation=200ms. (Providers are at liberty to measure more often than every 200ms and to report that fact)

- The 99th percentile value, i.e. $IPDV_{(99)}$, is chosen so that a stable value is achievable for the 1500 singleton IPDV values (5x60x5) obtained over the measurement period.
- The IPDV metric is reported in ms; Accurate to 1 ms, rounded up.
- The minimum reported one-way IPDV is 0 ms.
- One IPDV value is reported for each test period.

See Sections 3.6 and 4 for more discussion of the reporting and use of these metrics, and Appendix C for more discussion of IPDV measurement and reporting.

### 3.4.4 Packet loss ratio [PLR]

The definition of packet loss ratio (PLR) follows the approach defined in RFC 2680.

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 0 if the packet was received or 1 if the packet was not received. A packet is deemed to be lost if its one way delay exceeds an agreed time $T_{max}$. We draw on the ITU-T Y1540 provisional value of 3 secs as the recommended value for $T_{max,}$ when a packet is deemed to be lost. A packet is also counted as not received if it is corrupted in transit.

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 number of lost packets at the destination reference point divided by the number of packets sent at the sender reference point to that destination.

- The recommended maximum evaluation interval = 5 minutes
- Recommended mean packet separation=200ms. (Providers are at liberty to measure more often than every 200ms and to report that fact)
- PLR metric is reported as a percentage, accurate to 0.1 percent
- The minimum reported one-way PLR is 0.
- One PLR value is reported for each test period.

## *3.5 SLA definition for the "low latency" class*

Using the metrics defined in the previous section, it is possible to define the performance characteristics of the Low Latency service class. Details of the measurement approach to be taken are presented in Section 4. In this section we recommended performance characteristics for IP or MPLS traffic that may traverse the networks of multiple providers. The issue of how the total impairment budget is allocated among multiple providers is discussed below.

We draw on the service class definitions of Y.1541. Y.1541 defines two classes that are potentially suitable for VOIP (Classes 0 and 1). Class 0 is the more stringent, and where possible, providers should aim to deliver the SLA targets for class 0.

The parameters specified in Y.1541 for class 0 are as follows:

- IPTD: 100 msec (One Way Delay in IPPM terms)
- IPDV: 50 msec
- IPLR: $1 \times 10^{-3}$ (One Way Packet Loss in IPPM terms)

Where geographic distance prevents the delivery of class 0, class 1 may be provided. (We observe, however, that the upper bound on the mean IPTD for class 1 is likely to be unsuitable for interactive voice.)

The parameters specified in Y.1541 for class 1 are as follows:

- IPTD: 400 msec (One Way Delay in IPPM terms)
- IPDV: 50 msec (OWJ in IPPM terms)
- IPLR: $1 \times 10^{-3}$ (One Way Packet Loss in IPPM terms)

An evaluation interval of 5 minutes is suggested for IPTD, IPDV, and IPLR, and in all cases, the interval must be reported. Any 5 minute interval observed should meet these objectives.

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 24/7 excluding periods of unavailability and planned outages. See Section 7 for further discussion of maintenance windows.

## *3.6 Impairment Budgets*

### 3.6.1 Introduction

To support real-time traffic in multi-provider VPNs with the desired quality of service, the end-to-end impairment objectives for Low Latency class, as defined above, 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. Before suggesting a particular algorithm we look at requirements and consider different general approaches.

### 3.6.2 Requirements

Any algorithm for impairment budget apportionment 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 |

### 3.6.3 The challenge of budget apportionment and subsequent concatenation

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 major challenge for applications with stringent performance requirements.

Section 3.5 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 or allocate impairments in various parts of the provider's 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 approaches that could be taken in allocating total impairment targets among network segments can be characterized by the amount of information shared among segments and at what point in the design process and subsequent operation of the network that information must be gathered and assessed. Each approach has its pros and cons. Appendix A outlines various approaches that have been considered.

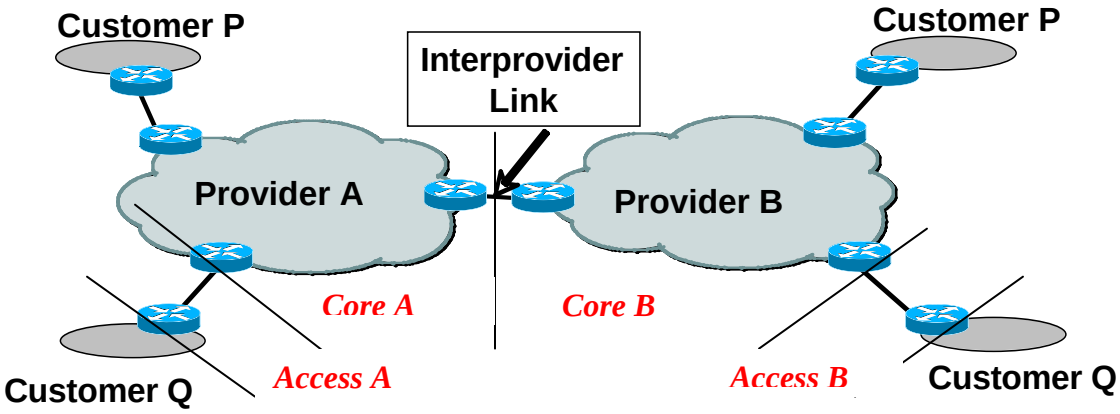
This paper proposes a fixed allocation of impairment budgets to each provider rather then apportionment on a path by path basis.

### 3.6.4 Consideration of Approaches to Impairment Allocation

Underlying delay does not rapidly change and can be numerically added to accurately derive the likely end to end outcome. However rapidly varying metrics such as delay variation cannot be so easily numerically concatenated. As part of the impairment budget allocation process therefore a pragmatic approach to setting budgets must factor in the statistical nature of these metrics while still, as much as reasonably possible, resulting in a high probability that the concatenation of any network sections will still meet the desired end to end outcomes.

A key consideration for the approach adopted in this paper is the requirement that any provider should not be reliant on a priori knowledge of the performance of other provider's network before being able to design his own network or prepare a commercial proposal for an end to end service. Guidelines are required that enable providers to design their networks in isolation, but at the same time have a high degree of confidence that end to end targets will be met when any reasonable concatenation of provider network segments are subsequently required to achieve an end to end service.

The reference model used is as follows:




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

### 3.6.5 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 customer's side of the first PE router.

For many networks, 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.

### 3.6.6 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 use of no more than three concatenated core network providers would be a reasonable maximum. In addition we assume one access provider per end, and zero or one metro providers per end, for a maximum of 7 concatenated networks. (It is possible that a single provider might divide his network into a metro network and a core network and allocate impairment budgets to each network separately.) As discussed in the following section, a larger amount of IPDV and PLR budget will be allocated to the access networks. Metro networks are not treated differently from core networks as far as budget allocation is concerned – that is, we allow two access networks and up to five "non-access" networks. In the remainder of this section we use "Core" to describe all "non-access" segments.

For any national (e.g. trans United States or Australia) or regional (e.g. Western Europe or Eastern Asia) service, the end to end target is ITU Y1541 Class 0. 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, but should be as close to Class 0 targets as practical. In either case, the end to end budgets should have a high probability of being met if

1. all providers consume their maximum impairment budget allocation for normal base line performance
2. The probability of more than one operator simultaneously operating in the upper range of the performance budget for varying impairments is very low, i.e. less than 0.1%.

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 or sales representatives cannot rely

on this when dimensioning or offering a network service. The following section proposes an approach to dimensioning and allocating impairment budgets among providers.

### 3.6.7 Budget allocation for planning purposes

This section outlines a simple, pragmatic solution for the apportionment of impairments. The most complex impairment metric to allocate is delay variation, discussed below in the following subsection. Impairment budget allocation methods for OWD and PLR are discussed in subsection 3.6.7.2.

Appendix B gives examples of the end to outcomes resulting from this allocation of impairment budgets. Appendix C describes an enhancement to the approach to IPDV concatenation and reporting.

#### 3.6.7.1 Budget allocation for IPDV and IPDV concatenation

The simple arithmetic division of delay variation budget across multiple providers would ordinarily result in a more stringent requirement than is actually required to achieve the end to end targets.

The method of determining budget allocation described in this section is less demanding on individual providers as it acknowledges the statistical nature of delay variation. The solution can also be used for the advanced signaled or accumulation approaches or it can be further refined to obtain tighter end-to-end performances values. The solution can be extended for more network segments if needed.

The approach presented here allocates a significant proportion of the IPDV impairment budget to each access segment, with each core segment having a lesser fixed budget. The approach also allocates a fixed IPDV budget for core network segments, irrespective of the number of core network segments in any resulting services.

Instead of requesting each provider to monitor a full IPDV distribution as would be required for the convolution method, each provider is requested to approximate the IPDV into three regions of magnitude:

- Low IPDV - normal conditions
- High IPDV but within bounds - unusual conditions handled by moderate buffers
- Extreme IPDV that exceeds bounds - extreme conditions where forecast has failed

More precisely, each provider commits to deliver a service with specified probability bounds of seeing a “Low IPDV” 5 minute interval and a “High IPDV” 5 minute interval (with the implied bound on the probability of seeing “Extreme IPDV"). Note that IPDV here is understood to mean the 99-th percentile of the singleton 5 minute measurements as discussed in Section 3.4.1.

The suggested thresholds defining Low, High and Extreme regions offered below are intended to be readily achievable for any Core provider offering a low latency class:

- Low IPDV < 2ms (normal case). Probability of any measurement interval with the IPDV being in this region is 0.99
- High IPDV: 2-6 ms. Probability of seeing a measurement interval with IPDV in this region is 0.00999
- Extreme IPDV > 6ms. Probability of being in this region is $\leq 1 \times 10^{-5}$

Similarly, the following thresholds are proposed for Access networks, where lower link speeds mandate more generous allocations of IPDV:

- Low IPDV < 16ms (normal case). Probability of any measurement interval being in this region is 0.99
- High IPDV: 16-20 ms. Probability of being in this region is 0.00999
- Extreme IPDV > 20ms. Probability of being in this region is $\leq 1 \times 10^{-5}$

We refer to the declaration of these regions and the corresponding probabilities as the "two-point promise". The "two-point promise" is the essence of the statistical IPDV impairment allocation method proposed in this section.

Once the two-point promise is thus specified, the probability of end-to-end IPDV being less than the specified target is approximated as the probability of seeing a combination of "Low" and "High" intervals such that the sum of the maximum IPDV thresholds specified in the corresponding two-point promise is less that that target. The intuition behind this is that if end to end traffic encounters three "Low IPDV" core segments (with the IPDV threshold of 2ms each) and two "High IPDV" core segments (with the IPDV threshold of 6 ms), then the end to end IPDV across the five network segments will be below 3x2+2x6=18 ms with high probability.

For example, if one is interested in approximating the probability of end to end IPDV across five core segments, each one of those declaring the 2-point promise as specified above for the core (or metro) segments, then one can perform the following computation:

$$\begin{aligned}
&\text{Prob(e2e IPDV} < 20 \text{ ms)} \approx \text{Prob(Sum of IPDV thresholds in encounted 5 min intervals} < 20 \text{ ms)} \\
&\geq \text{Prob(all 5 intervals are "Low IPDV")} \\
&+ \text{Prob(4 out of 5 intervals are Low IPDV and one is High IPDV)} \\
&+ \text{Prob(3 out of 5 intervals are Low IPDV and 2 are High IPDV)} \\
&= (0.99)^5 + \binom{5}{4}(0.99)^4 * 0.00999 + \binom{5}{3}(0.99)^3 (0.00999)^2 = 0.99994
\end{aligned}$$

Note that any other combination of low and high intervals in the five network segments yields the sum of the corresponding IPDV thresholds exceeding 20 ms. Similarly, if any network has an interval of extreme IPDV then there can be no assurance that and end-to-end IPDV bound is met.

The meaning of this computation is that if each of the five network segments declare the "two-point promise" with the thresholds as specified above, then the probability of end to end IPDV across the concatenation of these five networks exceeding the desired target of 20 ms is very small ($1-0.99994 < 10^{-4}$ ).

Note that the above computation does not yield a reliable *theoretical* bound on the end to end probability of IPDV. However, in practice it is a very good (and typically conservative) approximation of this probability. Appendix C provides a different version of the "two-point promise" that does deliver a theoretically provable bound on the e2e probability at the expense of reporting an additional point on the IPDV distribution above the 99-th percentile.

A similar calculation can be used to show that with two access segments and five core/metro segments using the figures proposed above, the end-to-end IPDV can be kept below 50ms with probability 0.9998.

For comparison, a similar calculation with two access segments and only three core network segments yields the probability of end-to-end IPDV below 50 ms being 0.99997

For access segments with a peak data rate of under 2 Mb/s that are also used to carry best effort traffic on the same access link as the low latency class traffic, packet fragmentation techniques need 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.

### 3.6.7.2 Impairment allocation budget for OWD and PLR

For allocation of PLR and OWD, more straightforward methodologies can be used, since these metrics can be considered to be additive. We propose that for PLR, each access networks be allocated $4 \times 10^{-4}$, and that each core or metro network be allocated a PLR of $10^{-5}$. This would allow an end to end PLR of $8.5 \times 10^{-4}$, within the limits for ITU class 0 and class 1. (See Section 4.4.1.1 for discussion of the issues related to reporting PLR.)

We propose that each access network be allocated 25ms of OWD. Core networks less than 1200km edge to edge are allocated 10ms of OWD. 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 this the following formula would apply;

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

The additional OWD 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.

# 4 QoS Measurement

The monitoring and troubleshooting of interprovider SLAs requires measurement of QoS-related information along the path between customer sites. Some agreement among co-operating providers on common approaches to measurement will simplify the tasks of SLA monitoring and troubleshooting. This section lays out the requirements for QoS measurement in the interprovider context and suggests some best practices.

## *4.1 QoS Measurement Requirements*

The measurement methodology, protocol and reporting must be capable of estimating at least the set of QoS metrics defined in section 3.4 (one way delay, one way loss, one way delay variation) of packets transmitted between specified measurement points. It should be possible to perform measurements on-demand or on a periodic, ongoing basis.

In this document we have defined all metrics to be one-way. Thus measurements should also be made one-way. Because this raises some practical challenges (e.g. clock synchronization) there may be occasions where two-way measurements will be made (and one-way metrics may be estimated from the two-way measurements). If this is the case it must be noted and reported.

Measurement probe packets should traverse as much as possible the same path as user packets having the same QoS service class. They should also be subject to the same QoS mechanisms in routers along the path, implying that the DSCP value of probe packets should be appropriately set for the QoS class to be measured.

The measurement approach should not significantly impact production traffic, either through excessive link load from measurement probes or as the result of load placed on routers by the measurement processes such as generating and responding to probes.

Measurements are generally made between two points in the network. Any of the points mentioned in Section 2.2 (PE, CE, ASBR) may be useful points for one end of a measurement. We also introduce the concept of a Measurement PoP, a PoP which is specifically designated as a suitable endpoint for certain measurements. This concept is discussed in more detail below.

The measurement methodology should 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. It should ideally be possible to make PE-PE or CE-CE measurements, even when the PEs or CEs are contained in or attached to the networks of different providers. (Note however that large amounts of PE-PE or CE-CE probing raise scalability issues.)

The measurement methodology should specify how the errors in measurements are treated, and how results are processed in terms of any statistical treatment of data.

Finally, the measurement methods and protocol must provide means to limit and detect attempts to tamper with or alter the QoS metric estimates.

### 4.1.1 Service Provider Measurement Agreements

One of the major challenges of interprovider measurement is that there are so many valid options. This document narrows the options so that measurements made across the networks of multiple providers could be compared and combined to create meaningful and reasonably accurate end-to-end measurements. To that end, we list here the set of things that SPs would need to agree upon in the measurement area.

SPs should agree upon the metrics defined in Section 3.4. The methodology for measurement of these metrics should define the size of measurement packets, the measurement protocol (e.g. OWAMP), the frequency of tests, and the distribution of probe packets (e.g. Poisson) in test series. Note that this document suggests values for all these parameters.

It should also be possible to make measurements from within the network of one provider to the ASBR of a neighboring provider. A provider may also designate a measurement PoP as a location that has specific capabilities for measurement. In these cases SPs should agree on the volume of the test traffic that they will generate into each others' networks.

SPs should publish enough information about the location of measurement devices that are available for customer or other SP-initiated measurements to enable customers or other SPs to make rational choices of where to direct their measurement traffic.

Co-operating providers should agree on the clock accuracy they will support. We propose a maximum error of 100μs for measurement devices in measurement PoPs, and a maximum error of 1ms for other measurement devices (e.g. CEs, PEs, or devices co-located with them.)

In order to support diagnostics and SLA conformance tracking, each provider must retain QoS measurement data for some agreed upon period.

## 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 a number of measurement methodologies are already in use. In this section we describe some of the options that exist within the realm of active (i.e. probe-based) measurement, as distinct from passive measurement in which the actual data traffic is monitored to gather performance data.

The sources and sinks of probes may be either dedicated measurement devices, routers that are dedicated to measurement tasks, or routers that support both data traffic and measurement probes. The location of measurement points may include:

- 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 or a matrix of such measurements among various points. It is also possible to report average measurements or other statistics computed over a number of different point-to-point measurements – such statistics clearly become less useful if the measurement points span widely different geographic areas.

When selecting measurement points, the goal is to capture the properties of the paths traversed by real customer traffic as much as possible. In general it will only be possible to approximate the path of customer traffic with a bounded number measurement devices. See Section 4.6 for further discussion of this issue.

To enable measurement of QoS parameters across multiple provider networks, one of the following methods could be used:

- Each provider agrees to use a common measurement protocol and to make probe points available to other providers, enabling measurements to be made along the end-to-end path
- Each provider network uses its own methods and probe devices to collect measurements on a per-provider basis, with these measurements being combined to estimate the concatenated end-to-end performance

Note that even the latter requires co-operation among interconnected SPs in terms of the protocol and availability of probe points to measure the QoS parameters of the inter-provider links.

## 4.3 QoS Measurement Protocols

ICMP-based PING measurement of TWPD, TWPL, and Instantaneous Bi-Directional Connectivity has historically been used by a number of providers when monitoring networks to deliver QoS-oriented SLAs. Vendor-proprietary measurement protocols have also been developed and used by some providers and end customers. In general, we recommend that, for inter-provider performance testing, open testing protocols should be used. In this document we propose that the IPPM protocol OWAMP [OWAMP] (or a protocol compatible with it) should be used for one-way measurements, with TWAMP [TWAMP] as an alternative if two-way measurements are to be used. (Note that this

document recommends one-way measurements but allows two-way as long as the distinction is reported.) In addition, there is an ongoing work that would allow the use of a lightweight version of TWAMP for one-way measurements. With this approach TWAMP and its simple version TWAM-lite can provide simple but reliable one-way and two-ways performance measurements. One of the possible avenues that needs to be explored is the use of OAM multihop protocols for inter-provider performance testing. Such an approach could reduce significantly the operational burden of network performance monitoring.

## *4.4 Reporting of Measurement results*

SPs need agree on the reporting methods. At the minimum there should be agreed processes for the exchange of hard copies of the performance results, including the content and format of such reports.

It is highly desirable that SPs agree on methods for the electronic exchange of measurement reports. Such an agreement would include both the content of the reports and a protocol for exchange of the reports.

The frequency of reports should be agreed upon. It would also need to be agreed whether reporting of QoS information among providers is a normal, ongoing activity or whether it is only triggered by requests (e.g. to troubleshoot a particular customer problem.) We propose that daily reporting would be ideal.

Reports should contain only aggregated data. Aggregated data should be available at different aggregated levels (by the fraction of an hour, by hour, daily, monthly – depending on the report) and statistics of the aggregates (mean, median, quantiles, number of measurements) should be reported.

The report should include at the minimum:

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

### 4.4.1 Proposal for reporting of measurement results

This section defines one method to report the set of QoS metrics defined in section 3.4 (PLR, OWD, one way IPDV) of packets transmitted between specified measurement points. As the rest of the paper, we will use VPN provider interconnection as our primary focus, but the intent is that the reporting is applicable in the broader Internet context as well.

As noted above, we propose to statically divide impairment budgets among the participating networks so that the budget per network segment can be checked. We recommend the three steps below to ensure that each provider can formulate an attractive

end-to-end SLA and also have the information necessary to trouble shoot for a VPN customer across multiple providers.

Each operator needs to measure the metrics as defined in Section 3.4 over 5 minute periods, 1500 samples per period:

- Loss (PLR)
- Delay (OWD)
- One-Way IPDV

As noted above, individual measurements will not be reported, but the appropriate 5 minute statistics (means for loss and delay, 99$^{th}$ percentile for IPDV) may be reported, as described below. Appendix C proposes an enhanced IPDV reporting methodology, which may be used in addition to the one described here.

#### 4.4.1.1 Monitoring and comparison to threshold

For each metric, the 5 minute measurements need to be monitored and compared with the threshold values in suggested in table 1. From a practical point of view it is an advantage to report (or act upon) as little as possible. Thus not all quantities need to be reported (or acted upon) for all classes. For the selection of important QoS parameters in the various traffic classes, see table 1.

| Result of 5 min Measurement | Low Latency Class Report |
|---|---|
| IPDV > 2ms (core/metro) | Report value |
| IPDV > 6 ms (core/metro) | Report value |
| IPDV > 16ms (access) | Report value |
| IPDV > 20ms (access) | Report value |
| PLR > $10^{-5}$ (core/metro) | Report value |
| PLR > 4 x $10^{-4}$ (access) | Report value |
| PLR > $10^{-2}$ | Report unavailable |
| OWD > 25 ms (access) | Report value |
| OWD > 10ms + distance allowance (core/metro) | Report value |

**Table 1 Thresholds for reporting among providers. Values below thresholds are not reported.**

Providers are only required to report (or act) when during a certain 5 minutes period a measured value is above the threshold. The suggested threshold value are set with such a margin that we normally don't need to report (or act upon) anything, but still not so large that the E2E budget is in danger when we are below this threshold. In practice we only report when a link or a cluster is very highly loaded and thus have problems with the QoS levels.

We note that a single 5-minute measurement interval with the measurement frequency recommended in Section 3.4.4 is not sufficient to accurately measure PLRs as low as these PLR thresholds. Thus, in order to verify that a particular provider is meeting the targets, it will be necessary to keep track of multiple 5-minute statistics. For example, to verify that a provider has delivered a loss rate of less than $10^{-5}$, at least 67 measurement intervals would be required to ensure enough samples have been taken at the recommended measurement frequency.[2]

#### 4.4.1.2 Reporting the measurement results:

With the threshold above providers need only to report to each other when something unusual occurs. We are working on the assumption that providers don't try to cheat each other on purpose but rather that operators report events that might endanger the E2E SLA, so that the owner of the E2E SLA can trouble shoot and constructively resolve problems.

The provider should specify the QoS problems for a relevant part of its domain in each case and not just report "problems anywhere in the domain" quite frequently. The provider should not report more to any provider than what is relevant to him. The reports could potentially therefore specify the VPNs which a certain measurement with a bad value will affect.

## *4.5 QoS Measurement Security Considerations*

Security is discussed in some detail in Section 6. Some specific security issues related to measurement involve the authentication of access and protecting the integrity of data. In particular:

- Integrity of measurement reports needs to be protected by standard cryptographic techniques.
- Authentication and access control mechanisms must be used to ensure that measurement reports are only made available to authorized parties.
- Access to a measurement probe devices, especially when access is permitted by other providers or customers, needs to be controlled by standard access control mechanisms.

## *4.6 Measurement considerations for VPN services*

When a VPN service spans the networks of multiple providers, there may be additional challenges in providing accurate end-to-end measurements for a given VPN customer. For example, it may be difficult for any one provider to determine the path that is taken by a particular VPN customer's traffic. And even if the path is known, it may be difficult to conduct measurements along that exact path, e.g. due to a lack of devices to respond to measurement probes at various points on the path.

[2] 67 intervals at 1500 probes per interval provides $10^5$ samples, so a single loss in those 67 intervals would represent the maximum allowable PLR.

The goals of the measurement techniques described above, therefore, are more modest than the delivery of precise performance data to a particular VPN customer. Instead, the primary goal is to allow a provider to make certain QoS assurances to a customer, knowing that

- the impairments that can be expected from other providers in the path, as described in Section 3.6.7, will enable those assurances to be met if all providers meet their impairment targets
- the reported measurements of each provider should indicate when a provider has failed to meet the targets.

# 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 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 [RFC4271] 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 an 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, delay variation, 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 may also 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 hierarchy 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" [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 re-advertised. 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 re-advertise 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 are 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:

a) Exchange QOS and Topology information in an opaque manner to enable service differentiation
b) Extend the protocol that follows current BGP configuration, policies and management via a backwards compatible technique
c) Enable BGP with fast convergence features for per service "SLAs".
   i. Announce multiple paths per prefix/service.
   ii. Withdraw multiple prefixes per AFI/SAFI/Topology/QOS class in one message.
d) Avoid interference with deployed features or availability mechanisms.
   i. Remove fate sharing of services.
   ii. 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 depends 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.

### 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 were to obtain access to the console port of a CE router). The CE device therefore cannot 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 [RFC4364]). 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.

As discussed in Section 2.3, packets must be marked correctly when crossing trust boundaries (CPI or PPI) in order to receive the appropriate service. Routers must therefore 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 cannot do so, then the service is subject to attack by simply flooding a router with too much traffic for it to examine.

Policing is also discussed in Section 3.2. 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 cannot 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 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 requested 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

The advent of interconnections where we undertake to deliver traffic with a specified quality brings new operational challenges. These are related to the operation of the differentiated services enabled interconnections, 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 QoS 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 for both the PPI and the CPI.

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 his own domain. Therefore it should not be a requirement for a provider to react to active probes (e.g. traceroute and ping) other than on the PE and ASBR nodes (although, as noted above, this capability may be made available selectively with appropriate authorization).

In the event of lost connectivity, service availability will depend on the efficiency of rerouting traffic. Each provider is responsible for rerouting the traffic within his domain and slow convergence will impact the SLA. This means that there is a direct connection between the requirement on fast rerouting of traffic and the formulation of the SLA metrics. Note that there is no need for any exchange of information on internal routing protocol rerouting performance.

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

## 7.2 Configuration and Maintenance

Due to the higher demands on performance (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 that may be involved in the interprovider QOS case, it is not possible to schedule a regular service window that is suitable to 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.

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 local 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). 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 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.

## 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.

We note that most PPI links carry aggregated traffic from many end customers and do not readily allow traffic from specific customers to be identified. (Option A interconnects are the exception to this). Thus it seems unlikely that accounting on a per-end-user basis can be achieved in many cases.

## 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 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 References

[AGG] Chan, K., Babiarz, J. and F. Baker, “Aggregation of DiffServ Service Classes”, Work in Progress, draft-chan-tsvwg-diffserv-class-aggr-03.txt, January 2006.

[RFC 4364] E.Rosen, Y.Rehkter, "BGP/MPLS IP VPNs", RFC 4364, February 2006.

[RFC 2475] 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

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

[CLASSES] Chan, K., Babiarz, J. and F. Baker, “Configuration Guidelines for DiffServ Service Classes”, RFC 4594, http://www.rfc-editor.org/rfc/rfc4594.txt, August 2006.

[RFC 3393] Demichellis, C. Chimento, P. IP Packet Delay Variation Metric for IP Performance Metrics (IPPM). RFC 3393, November 2002

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

[RFC 3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", April 2001.

[OPsecPractices] Merike Kaeo,"Operational Security Current Practices," draft-ietf-opsec-current-practices-03, May 2006.

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

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

[RFC 2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option," August 1998

[RFC 3246] Davie, B. et al. "An Expedited Forwarding PHB (Per-Hop Behavior)," March 2002.

[RFC 3552] Rescorla, E., and B. Korver, "Guidelines for Writing RFC Text on Security Considerations," July 2003.

[RFC 4111] Luyuan Fang, "Security Framework for Provider-Provisioned Virtual Private Networks (PPVPNs)", July 2005.

[RFC4271] Y. Rekhter et al. *A Border Gateway Protocol 4 (BGP-4).* January 2006.

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

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

[OWAMP] Shalunov, S., et al. A One-way Active Measurement Protocol (OWAMP), draft-ietf-ippm-owdp-16.txt, February 2006.

[TWAMP] Babiarz, J. et al. A Two-way Active Measurement Protocol (TWAMP). draft-ietf-ippm-twamp-00.txt, November, 2005.

[SecurityEfforts] C. Lonvick and D. Spak, "Security Best Practices Efforts and Documents," draft-ietf-opsec-efforts-03.txt, April 2006.

[Y.1541] Network Performance Objectives for IP-based Services, version 6, Oct 2005

[G.FEPO] Draft Framework for Achieving End-to-end Performance Objectives) Jan 2006

[Multisession BGP] draft-ietf-idr-bgp-multisession-02.txt, Chandra Appanna, John G. Scudder, March 2006.

# Appendix A. Discussion on impairment allocation approaches

For all approaches to impairment allocation, a "top-down" or "bottom-up" method could be applied. That is, percentages of the aggregated target (top-down) or fixed/negotiated values for impairments (bottom-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, in the near-term, a simpler approach is recommended. The recommendation should ultimately include an evolution path to more complex approaches.

| 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<br><br>Impairment allocation is formulaic among User, Access, Transit, and Peering segments | Information required is<br>a) type of link,<br>b) traffic service class and,<br>c) transit distance | No information is required to be shared among segments.<br><br>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<br><br>Paths having more that the assumed number of segments are not covered<br><br>Negotiation not supported. |

| | | | | |
|---|---|---|---|---|
| **Pseudo-static**<br>- some information is required to be shared among segments | The exact number of transit providers is determined<br><br>Impairment allocation is formulaic among User, Access, Transit, and Peering segments | Information required is<br>a) type of link,<br>b) traffic service class and,<br>c) transit distance<br>d) destination address<br>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<br><br>Negotiation not supported |
| **Signaled** (least simple/most flexible)<br>- 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<br><br>Impairment apportionment may be negotiated among segments and with Users | Information required is<br>a) type of link,<br>b) traffic service class<br>c) destination address<br>d) BGP tables, or other means to determine path or paths at the operator-level,<br>e) Network edge-edge performance information<br>Additional information that may be required includes<br>f) transit distance | Negotiation is supported allowing highly flexible apportionment among segments.<br><br>No predefined allocations are required.<br><br>Transit distance may not be required<br><br>Able to address cases where the objective can not be met by consulting user for relaxed objective<br><br>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.<br><br>Signaling may be required to negotiate with User when the requested objective cannot be met<br><br>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 2 – Summary of performance impairment apportionment approaches

# Appendix B. Examples of the application of budget allocations

In this appendix we 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.

Note that the allocation of IPDV in these examples uses the "low IPDV" thresholds from Section 3.6.7, and the arithmetic sum of those thresholds is shown just for illustrative purposes. Refer to Section 3.6.7 for more complete details of IPDV allocation.

## B.1 Case 1: Three Core Providers

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.

| | Link air path distance | IPTD budget (base) | Additional IPTD for Distance | Total IPTD | IPDV (low threshold) | IPLR |
|---|---|---|---|---|---|---|
| Access provider 1 | | 25ms | | 25ms | 16ms | $4 \times 10^{-4}$ |
| Core provider A | 300km | 10ms | 0 | 10ms | 2 | $1 \times 10^{-5}$ |
| Core provider B | 3000km | 10ms | 12 ms | 22ms | 2 | $1 \times 10^{-5}$ |
| Core provider C | 700km | 10ms | 0 | 10ms | 2 | $1 \times 10^{-5}$ |
| Access provider 2 | | 25ms | | 25ms | 16ms | $4 \times 10^{-4}$ |
| **Total CE to CE** | **4000km** | | | **92ms** | **38ms** | **$8.3 \times 10^{-4}$** |

Note that this meets the targets for ITU Class 0.

### B.2 Case 2: Transcontinental Service, 5 Core Providers

| | Link air path distance | IPTD budget (base) | Additional IPTD for Distance | Total IPTD | IPDV (Low Threshold} | IPLR |
|---|---|---|---|---|---|---|
| Access provider 1 | | 25ms | | 25ms | 16 ms | $4 \times 10^{-4}$ |
| Core provider A | 300km | 10ms | 0 ms | 10ms | 2 ms | $1 \times 10^{-5}$ |
| Core provider B | 3000km | 10ms | 12 ms | 22ms | 2 ms | $1 \times 10^{-5}$ |
| Core provider C | 10,000km | 10ms | 55 ms | 65 ms | 2 ms | $1 \times 10^{-5}$ |
| Core provider D | 2,000Km | 10ms | 5 ms | 15 ms | 2 ms | $1 \times 10^{-5}$ |
| Core provider E | 400Km | 10ms | 0 ms | 10 ms | 2 ms | $1 \times 10^{-5}$ |
| Access provider 2 | | 25ms | | 25ms | 16 ms | $4 \times 10^{-4}$ |
| **Total CE to CE** | **17,000km** | | | **172 ms** | **42 ms** | **$8.5 \times 10^{-4}$** |

Note that this meets the targets for ITU Class 1. Core providers A and E might be considered "metro" providers in this example.

# Appendix C. Alternative IPDV Concatenation Approach

The following is an alternative to the IPDV concatenation approach is described in Section 3.6.7.1. The analysis here is for a two-point promise for core/metro segments where "medium" IPDV is 2-6 ms and "high" IPDV is above 6ms. A similar analysis would hold for access networks with different points of medium and high IPDV.

## *C.1 What is promised by each provider*

In this proposal the “normal”, “medium” and “unusually high” regions are associated directly with the cumulative distribution function of the delay variation (rather than the 99-th percentile of delay variation). That is,

**Each provider declares that with the probability $p_{normal}$ the queuing delay in this provider’s network is $D_{normal}$ or less, and with the probability of $p_{medium}$ or more, the queuing delay falls in the range between $D_{normal}$ and $D_{medium}$. The remaining probability of seeing delay greater than $D_{medium}$ is then less than 1- $p_{normal}$ - $p_{medium}$.**

Note that the timescale of these declarations is not explicitly defined at this point, as the meaningful timescale depends on the choice of probabilities $p_{normal}$, $p_{medium}$ and $p_{high}$ (e.g.

if $p_{high}$ is very small, it can only be reliably estimated/verified at sufficiently large timescales).[3]

## *C.2 What end-to-end statements can be made*

In this approach, since the operators declare explicitly two points on the CDF of delay variation, one can compute an explicit lower bound on the end-to-end probability of delay exceeding a given threshold, without any further assumptions on the distribution of delay variation.

Specifically, assuming the independence of the delays[4], one can state that

Prob(e2e delay variation $\leq$ 18ms) $\geq$

Prob(delay in each of the 5 networks $\leq$ 2ms)+

Prob(delay in 4 of the 5 networks $\leq$ 2ms, and delay in one other is between 2 and 6 ms)+

Prob(delay in 3 of the 5 networks $\leq$ 2ms, and delay in two is between 2 and 6 ms)

This is a simple combinatorial problem yielding the result that

Prob(e2e delay variation $\leq$ 18ms) $\geq 0.9999$

This statement is true for any distributions satisfying the two points declared by each operator (under the independence assumption).

## *C.3 Verification of whether the promise is being delivered*

Conceptually one needs to verify per-packet probability of queuing delay falling into the specified time bounds. So the task consists of two steps (a) and (b) below:

a) Verifying that the experimental per-packet probability of not exceeding delay 2 ms being at least 0.99 is equivalent to verifying that the 99-th percentile of the sample is at most 2 ms. Therefore, verifying the declaration of probability being of queuing delay being less than 2 ms with probability of at least 0.99 can be done by simply reporting the 99-th percentile delay and comparing it with the desired statement. 1500 packets are sufficient to do so with reasonable confidence.

---

[3] Clearly, this dependence on the timescales that are potentially larger than the convenient five minute measurement interval is a drawback of this proposal. However, verification at larger timescales can be done at a relatively low cost, while the main advantage is the ability to provide a reliable end-to-end bound on delay variation independent of the distribution of the delay variation (which is not in general possible with the proposal in Section 3.6.7).

[4] In general the independence assumption may not hold

b) Verifying that per-packet probability of queuing delay being between 2 and 6 ms being at least 0.01 - 10^(5) is equivalent to verifying that the 99.999-th percentile of the sample is at most 6ms. However, the 99.999 percentile cannot be reliably verified on 1500 packet samples, and can only be verified at longer timescales (over sufficiently large number N of the 5 min intervals). To do such verification without actually keeping 1500 times N packet values, one can report the raw count of packets exceeding queuing delay of 6 ms, and keep the cumulative count of such per-interval count over N intervals. Then at the end of N intervals, one can estimate whether the experimental probability of queuing delay greater than 6 ms is within the expected probability range by comparing the experimental frequency of the packets with large delay to the promised probability values.

Both of these tasks can be accomplished by the standard hypothesis testing approach.

## *C.4 What is reported*

As described in the previous section, the approach requires that every 5 min intervals the two values are reported:

- The 99-th percentile of IPDV (the 5 minute IPDV statistic described in Section 3.4.3)
- The counter of the number of samples whose IPDV exceeds 6 ms

One can optimize the reporting by not reporting any values in an interval for which the 99-th percentile is below 2 ms AND there are no packets exceeding the queuing delay of 6 ms. This optimization is merely a reporting convention.

## *C.5 Comparison with the approach of section 3.6.7.1*

The key tradeoffs between the two proposals is between the amount of reporting, the guarantees that could be provided and how straight-forward trouble-shooting is. The proposal in this appendix provides a stricter end-to-end guarantee on the probability of not exceeding a given queuing delay, but the guarantee comes at the expense of reporting one extra counter per 5 min interval.

The table below summarizes all of the above.

| | *What is promised* | *What is reported over 5 min* | *What e2e Statements are made* | *Trouble shooting* |
|---|---|---|---|---|
| Section 3.6.7.1 | Two specific points on the CDF of the distribution of the 5-min | The 99-th percentile of OWD-OWD(min) | Claim of conservative estimate of the e2e probability of 99-th | All reported 5 minutes 99 percentile values may be directly used for trouble shooting, but appropriately large number of 5 min |

| | | | | |
|---|---|---|---|---|
| | 99-th percentiles | | percentile not exceeding specific queuing delay. No "guaranteed" bound on this probability in general | intervals are needed to reliably verify conformance to the two-point promise on the probability of seeing 5 min intervals with the normal, medium or high 99-th percentile. |
| Alternative Approach | Two specific points on the CDF of queuing delay distribution | The 99-th percentile of OWD-OWD(min) AND the number of packets experiencing very large delay | Guaranteed bound on the end-to-end probability of queuing delay not exceeding a specific value | The 5 minutes 99 percentile values can be directly used for trouble shooting. The 99.999 percentile values *may* be an indication of a problem, but requires caution not to trigger unnecessary alarms as longer timescale is needed for verification. |