BITAG

Broadband Internet Technical Advisory Group

Real-time Network Management of Internet Congestion

A BROADBAND INTERNET TECHNICAL ADVISORY GROUP
TECHNICAL WORKING GROUP REPORT

A Uniform Agreement Report

Issued:
October 2013

Copyright / Legal Notice

This document and the information contained herein is provided on an "AS IS" basis and BITAG AND THE CONTRIBUTORS TO THIS REPORT MAKE NO (AND HEREBY EXPRESSLY DISCLAIM ANY) WARRANTIES (EXPRESS, IMPLIED OR OTHERWISE), INCLUDING IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, FITNESS FOR A PARTICULAR PURPOSE, OR TITLE, RELATED TO THIS REPORT, AND THE ENTIRE RISK OF RELYING UPON THIS REPORT OR IMPLEMENTING OR USING THE TECHNOLOGY DESCRIBED IN THIS REPORT IS ASSUMED BY THE USER OR IMPLEMENTER.

The information contained in this Report was made available from contributions from various sources, including members of Broadband Internet Technical Advisory Group, Inc.'s Technical Working Group and others. Broadband Internet Technical Advisory Group, Inc. takes no position regarding the validity or scope of any intellectual property rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this Report or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

About the BITAG

The Broadband Internet Technical Advisory Group (BITAG) is a non-profit, multi-stakeholder organization focused on bringing together engineers and technologists in a Technical Working Group (TWG) to develop consensus on broadband network management practices and other related technical issues that can affect users' Internet experience, including the impact to and from applications, content and devices that utilize the Internet.

The BITAG's mission includes: (a) educating policymakers on such technical issues; (b) addressing specific technical matters in an effort to minimize related policy disputes; and (c) serving as a sounding board for new ideas and network management practices. Specific TWG functions also may include: (i) identifying "best practices" by broadband providers and other entities; (ii) interpreting and applying "safe harbor" practices; (iii) otherwise providing technical guidance to industry and to the public; and/or (iv) issuing advisory opinions on the technical issues germane to the TWG's mission that may underlie disputes concerning broadband network management practices.

The BITAG Technical Working Group and its individual Committees make decisions through a consensus process, with the corresponding levels of agreement represented on the cover of each report. Each TWG Representative works towards achieving consensus around recommendations their respective organizations support, although even at the highest level of agreement, BITAG consensus does not require that all TWG member organizations agree with each and every sentence of a document. The Chair of each TWG Committee determines if consensus has been reached. In the case there is disagreement within a Committee as to whether there is consensus, BITAG has a voting process with which various levels of agreement may be more formally achieved and indicated. For more information please see the BITAG Technical Working Group Manual, available on the BITAG website at www.bitag.org.

BITAG TWG reports focus primarily on technical issues. While the reports may touch on a broad range of questions associated with a particular network management practice, the reports are not intended to address or analyze in a comprehensive fashion the economic, legal, regulatory or public policy issues that the practice may raise.

BITAG welcomes public comment. Please feel free to submit comments in writing via email at comments@bitag.org.

Executive Summary

The Internet, as is the case with many other networks such as highways and electricity grids, operates under the assumption that capacity will be set to a level such that total peak demand will occasionally exceed capacity. Further, the Internet is designed so that multiple users may dynamically share capacity and multiple services may share the same network links and routers, which is more efficient than offering individual users dedicated capacity or different services using separate links and routers.

Every link and router in the various networks that make up the Internet has a limit on its capacity to handle data. The capacity of each link and router in individual networks is determined by the equipment installed by the entity that runs each network in an attempt to optimize performance and cost; the lower the capacity relative to expected demand, the greater the probability that demand upon that link or router at times may exceed its capacity.

Significantly, a user's instantaneous demand for broadband Internet is bursty, meaning that it changes rapidly in time – and when aggregate instantaneous demand exceeds capacity on a network it causes congestion, which can degrade performance.

Network operators typically estimate demand months to years in advance, and use such demand estimates to plan a schedule for capacity upgrades. Since it may take months to implement a capacity upgrade, the time scale for managing congestion in this manner is months to years. Thus, although capacity planning can greatly affect how much congestion occurs on a network over time, it cannot react to congestion as it occurs.

The impact of congestion upon applications depends on the duration of congestion – which can vary from thousandths of a second up to hours or more – and the nature and design of the application. If the duration of congestion is short enough or the application is tolerant enough of congestion, a user will not notice any degradation in performance. Congestion is thus a problem only when its duration is long enough to be disruptive to applications. Congestion in a network can occur for a wide variety of reasons, some of which can be anticipated and some of which cannot.

This report describes how network resources are allocated on a short time scale in order to, among other objectives, manage congestion on the network, and how such congestion management impacts applications and users.

Congestion management practices are an important subset of network management practices implemented by a variety of parties or organizations, including Internet Service Providers (ISPs) and Application Service Providers (ASPs). Policymakers have expressed great interest in learning what congestion management practices are used in the Internet and how these practices impact users and the broader Internet ecosystem. Furthermore, an understanding of congestion management techniques and practices is crucial in discussions about reasonable network management.

One of the key design questions about any congestion management practice concerns the subset of network traffic to which the practice is applied, and its impact upon users and applications. Network operators apply some practices to all traffic on their networks, whereas in other cases practices are applied only to the traffic of specific users or to the traffic associated with specific applications. Application- or user- based congestion management practices may achieve better performance for selected applications. They also may enable service providers to offer connectivity products that cater to particular customer's tastes or needs. However, they add complexity, which may result in added costs that each network operator will evaluate. In some cases application- or user- based congestion management practices may be harmful to applications.

Congestion management practices are composed of generic technical building blocks, described in this report as traffic management "techniques". This report discusses a range of user- and application- based congestion management techniques, including classification of packets, reservation of resources for particular network flows, storage of content in multiple locations, rate control, routing and traffic engineering, packet dropping, and packet scheduling.

Congestion management techniques may be combined to offer a collection of capabilities in various network architectures, and can create services with differentiated performance either within an operator's network or end-to-end. There are also architecture-specific implementations of congestion management techniques for broadband Internet access over cable, telephone, and cellular networks and for Content Delivery Networks. The offerings of a service provider often include multiple services that may utilize the same network links and routers. While there are benefits and efficiencies to sharing capacity between multiple services, such sharing of capacity also requires the use of congestion management practices.

Congestion management "practices" are the uses of particular techniques by particular network operators to avoid, limit, or manage congestion. This report illustrates a range of congestion management practices that show how providers may combine user- or application- based congestion management techniques, including traffic shaping, prioritization, transcoding, resource reservation, and preferential treatment.

The report begins in Sections 1 and 2 by giving an overview of congestion and BITAG's interest in the issue. Section 3 defines congestion and describes instances in which congestion can occur, the locations in the network where congestion can occur, the indicators of congestion, and the impact congestion can have on applications.

In Section 4, the report articulates the differences between congestion management techniques and congestion management practices, and describes the different time scales at which congestion can be seen to occur in the network. This section also describes the parties that implement congestion management practices and on what basis.

Although all congestion management is important, in order to limit scope and length Sections 5-7 focus on congestion management techniques and practices that: (1) are implemented or potentially implemented in a network that supports consumer broadband Internet access services; (2) act on a time scale of minutes or less; (3) are used for purposes of congestion management; and (4) are based on user or application.

In Section 5, the report focuses on specific congestion management techniques. Section 6 gives specific examples of congestion management practices that are based on user or application. Finally, Section 7 gives the Technical Working Group's recommendations.

At a high level, the recommendations of BITAG's Technical Working Group are:

- **ISPs and ASPs should disclose information about their user- or application-based network management and congestion management practices for Internet services in a manner that is readily accessible to the general public**. This information should be made available on network operators' public web sites and through other typically used communications and channels, including mobile apps, contract language, or email. ISPs and ASPs may choose to use a layered notice approach, using a simple, concise disclosure that includes key details of interest to consumers complemented by a more thorough and detailed disclosure for use by more sophisticated users, application developers, and other interested parties. The detailed disclosure should include: descriptions of the practices; the purposes served by the practices; the types of traffic subject to the practices; the practices' likely effects on end users' experiences; the triggers that activate the use of the practices; the approximate times at which the practices are used; and which subset of users may be affected. The disclosures should also include the predictable impact, if any, of a user's other subscribed network services on the performance and capacity of that user's broadband Internet access services during times of congestion, where applicable.

- **Network operators should use accepted industry "Best Practices," standardized practices, or seek industry review of practices.** Network standards setting organizations and technical industry bodies produce considered recommendations of Best Practices and standard practices for a variety of operational issues including congestion and congestion management. Where network operators see the need for an innovative solution that has not been standardized or documented as a Best Practice, these network operators should consider bringing their unique network or congestion management practices to such groups for discussion and documentation.

- **When engaging in a congestion management practice that could have a detrimental impact on the traffic of certain users or certain applications, the practice should be designed to minimize that impact.** Some congestion management practices may cause certain users or certain applications to experience performance degradation. ISPs and ASPs should seek to minimize such degradation to the extent possible while still managing the effects of the congestion that originally triggered the use of the practice.

- **If application-based congestion management practices are used, those based on a user's expressed preferences are preferred over those that are not.** User- and application- agnostic congestion management practices are useful in a wide variety of situations, and may be sufficient to accommodate the congestion management needs of network operators in the majority of situations. However, at times network operators may choose to use application-based congestion management practices, in which case those that prioritize application traffic according to a user's expressed preferences are preferred over those that do not.

- **If application-based criteria are used by a network operator, they should be tested prior to deployment and on an ongoing basis.** Application-based classification by network operators (e.g., using deep packet inspection) can sometimes be erroneous. If network operators choose to use application-based criteria for congestion management, the accuracy of the classifier should be tested before deployment.

- **ASPs and CDNs should implement efficient and adaptive network resource management practices.** ASPs and CDNs should match use of network resources to the performance requirements of the application. Applications should be designed to efficiently and adaptively use network resources, to the extent feasible given the application's requirements.

Table of Contents

1. Issue Overview

The Internet, as is the case with many other networks such as highways and electricity grids, operates under the assumption that capacity will be set to a level such that total peak *demand* will occasionally exceed *capacity*. Further, the Internet is designed so that multiple users may dynamically share capacity and multiple services may share the same network links and routers, which is more efficient than offering individual users dedicated capacity or different services using separate links and routers.

Every link and router in the various networks that make up the Internet has a limit on its capacity to handle data. The capacity of each link and router in individual networks is determined by the equipment installed by the entity that runs each network in an attempt to optimize performance and cost; the lower the capacity relative to expected demand, the greater the probability that demand upon that link or router at times may exceed its capacity.

Significantly, a user's instantaneous demand for broadband Internet is *bursty*, meaning that it changes rapidly in time – and when aggregate instantaneous demand exceeds capacity on a network it causes *congestion*, which can degrade performance.

Network operators typically estimate demand months to years in advance, and use such demand estimates to plan a schedule for capacity upgrades. Since it may take months to implement a capacity upgrade, the time scale for managing congestion in this manner is months to years. Thus, although capacity planning can greatly affect how much congestion occurs on a network over time, it cannot react to congestion as it occurs.

The impact of congestion upon applications depends on the duration of congestion – which can vary from thousandths of a second up to hours or more – and the nature and design of the application. If the duration of congestion is short enough or the application is tolerant enough of congestion, a user will not notice any degradation in performance. Congestion is thus a problem only when its duration is long enough to be disruptive to applications. Congestion in a network can occur for a wide variety of reasons, some of which can be anticipated and some of which cannot.

This report describes how network resources are allocated on a short time scale in order to, among other objectives, manage congestion on the network, and how such congestion management impacts applications and users.

One of the key design questions about any congestion management practice concerns the subset of network traffic to which the practice is applied, and its impact upon users and applications. Network operators apply some practices to all traffic on their networks, whereas in other cases practices are applied only to the traffic of a subset of specific users, to a subset of types of applications, to all instances or specific instances of applications, or to specific components of such applications. Application- or user- based congestion management practices may achieve better performance for various application-related traffic. Such practices also may enable service providers to offer connectivity products that cater to particular customer's tastes or needs. However, they add complexity, which may result in added costs that each network operator will evaluate. In some cases application- or user- based congestion management practices may be harmful to applications.

This report focuses on real-time Internet traffic management practices based on users or applications that are used on networks operated by Internet Service Providers (ISPs) and Application Service Providers (ASPs) (known as "network operators" throughout this report) for the purposes of congestion management.[1] Network management practices used by network operators for purposes other than congestion management are outside the scope of this report. Practices that are not implemented in real-time are also outside the scope, including usage caps and usage charges.

The analysis distinguishes between traffic management "techniques," which are generic technical building blocks, and traffic management "practices," which are the applications of particular techniques by particular network operators to avoid, limit, or manage congestion. With respect to techniques, the analysis considers where in the network, and at which layer, a traffic management technique is applied, and what type of traffic management functionality is applied. With respect to practices, the analysis considers who decides whether a traffic management practice is applied and on what basis. It is important to examine the criteria and indicators of congestion that trigger a practice.

2. BITAG Interest in the Issue

Congestion management practices are an important subset of network management practices implemented by a variety of parties or organizations, including Internet Service Providers (ISPs) and Application Service Providers (ASPs). Policymakers have expressed great interest in learning what congestion management practices are used in the Internet and how these practices impact users and the broader Internet ecosystem [FCC 07-31].

Policymakers often comment that network architectures and technologies may impact congestion management practices, but are looking for guidance as to how this occurs. Furthermore, an understanding of congestion management techniques and practices is crucial in discussions about reasonable network management.

[1] For purposes of this report, an Internet Service Provider (ISP) is defined as a provider of broadband Internet access service, an Application Service Provider (ASP) is defined as a provider of applications used on broadband Internet access services, and a network operator is defined as an ISP, or an ASP that operates a network. Some ASPs operate networks that interconnect with ISPs, while other ASPs attach servers directly to ISPs.

3. Characterization of Congestion

It is important to understand what is meant by the term “congestion” in the context of the Internet, and this section of the report provides an overview. Section 3.1 discusses demand, capacity, and congestion. Section 3.2 describes in what instances congestion can occur. Section 3.3 describes the locations in the network where congestion can occur. Section 3.4 describes the indicators of congestion. Section 3.5 describes the impact congestion can have on applications.

3.1. Definition of congestion

As pictured in Figure 1, a user’s instantaneous *demand* for broadband Internet (measured in bits per second) is *bursty*, meaning that instantaneous demand changes rapidly in time.[2] A user’s average demand, measured over several days, is much lower than the user’s peak demand. The Internet is designed so that multiple users may dynamically share *capacity*, a concept called *statistical multiplexing*.

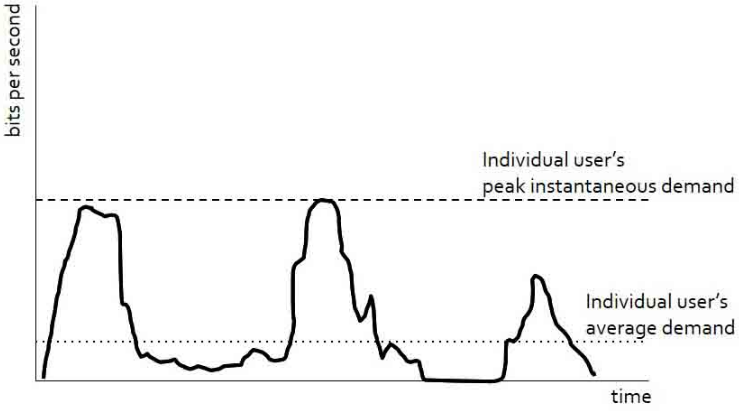

Figure 1. A single user’s demand for broadband Internet, measured in bits per second.

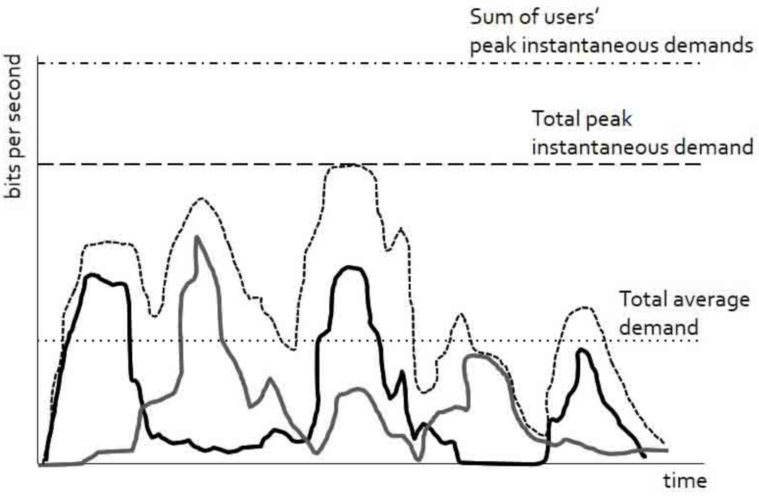

Figure 2. The sum of two users' demand for broadband Internet.

Figure 2 illustrates two users sharing capacity. One user’s instantaneous demand is shown as a solid black line, another user’s instantaneous demand as a solid grey line, and the sum of their instantaneous demands as a dashed black line. The total average demand is simply the sum of the user’s individual average demands. However, users’ individual instantaneous demands are usually uncorrelated with each other, so that they burst to high levels at different times. As a result, the total peak demand (the highest point on the dashed curve) is usually far less than the sum of users’ individual peak

[2] Figures 1-3 are for illustrative purposes only, and do not represent actual measured network data.

demands (the dash dot line at the top of the figure) due to the fact that users' individual peaks are typically non-concurrent. Some of the reasons for occurrences of high demand and of fluctuation in demand, along with the potential duration of such demand, are described in Section 3.2.

All links and routers in a network have a limit on their capacity to handle data, as described in Section 3.3. As illustrated in Figure 3, for purposes of this report, *congestion* is defined as the effect upon network performance during time periods in which instantaneous demand exceeds capacity.[3]

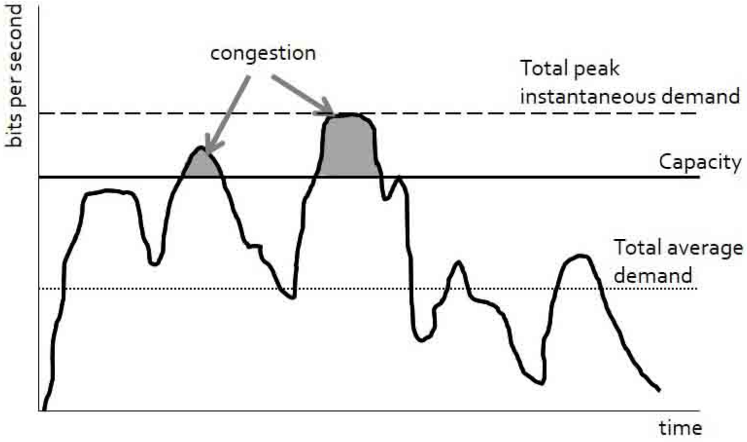

Figure 3. Congestion occurs when instantaneous demand exceeds capacity.

The Internet operates under the assumption that capacity will be set to a level such that total instantaneous peak demand will occasionally exceed capacity, see e.g. [Kurose and Ross, section 3.1]. This design is based upon the cost efficiency that can be gained through dynamic sharing of capacity. The capacity of each link and router in a network is determined by the equipment installed by the entity that runs the network in an attempt to optimize performance and cost; the lower the capacity relative to expected to demand, the greater the probability that instantaneous demand upon that link or router may at times exceed its capacity. Because not all users are active or fully use their maximum Internet connection speed at the same time, a network operator may install capacity in a link or router at a level above the total average demand but below the total instantaneous peak demand. This well-established practice of network design lowers the cost of creating a network and providing connectivity. It is used not only in the Internet, but also on highways, electricity networks, and air transportation networks, since it would be prohibitively expensive to add enough capacity to ensure that congestion never occurs.

Congestion may cause an increase in the *end-to-end packet delay*, which is the delay from the time a packet is transmitted by the source until it is received by the destination. Congestion may also cause an increase in *end-to-end packet loss,* which is the proportion of packets that do not arrive at the destination. These indicators of congestion and methods for network operators to measure congestion are discussed in Section 3.4.

[3] Alternate definitions of *congestion* include the effect upon network performance during time periods when (1) average demand exceeds capacity over a specified measurement interval, (2) the load over a specified measurement interval exceeds a specified threshold, and (3) packets are dropped by a router [Evolution of Internet Congestion]. These indicators of congestion are discussed in Section 3.4.

The duration of congestion can vary from milliseconds (thousandths of a second) to hours. The impact of congestion upon applications depends on the duration and severity of congestion and the nature and design of the application. If the duration of congestion is short enough or the application is tolerant enough of congestion, a user will not notice any degradation in performance. Congestion is thus a problem only when its duration is long enough to be disruptive to applications. The impact that congestion has on users, applications, and ASPs is discussed in Section 3.5.

The total average demand across an operator's network varies by hour and day of the week. For consumer wireline networks, average demand is usually highest during the evening hours of each day, typically exhibiting a pattern similar to that illustrated in Figure 4.

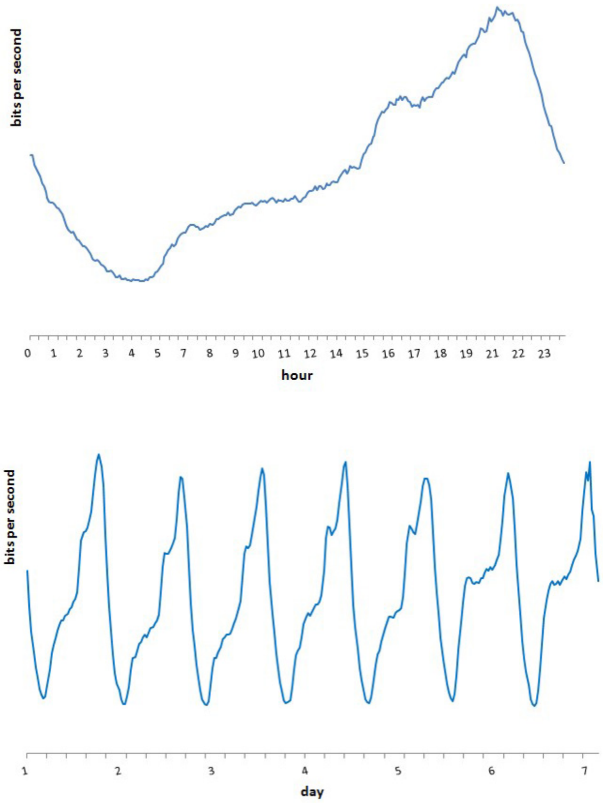

Figure 4. Total average demand by hour and day of week.

If capacity sufficiently exceeds the total average demand during the busiest hours, then the duration of congestion is generally short – milliseconds to seconds. This is the desired situation, as most users and applications will not experience a reduction in perceived performance. Occasional short-term congestion is unavoidable.

In contrast, if capacity does not sufficiently exceed the total average demand during the busiest hours, then the duration of congestion may be longer – minutes to hours – which will significantly degrade the perceived performance of most users and applications. In this situation, the only effective solutions to long-term congestion are to either increase capacity or decrease demand (which are not core topics discussed further in this document). A network operator will usually schedule upgrades to the links and routers to increase capacity months before it predicts that such long-term congestion will occur [RFC 6057]. The cost of adding capacity varies according to the technology, and is generally the highest in access networks (which include the portions of the network often referred to as the "last mile"). A network operator will consider cost and performance using each particular access technology when deciding how and when to increase capacity. Reductions in average demand can, among other ways, be accomplished by creating more bandwidth-efficient applications and services or altering users' incentives through pricing plans.

3.2. Occurrence and Duration of Congestion

Congestion in an ISP or ASP network can occur for a wide variety of reasons, some of which can be anticipated and some of which cannot. For purposes of this report, the causes of congestion can roughly be classified as: recurrent congestion, predictable events, unpredictable events, and random congestion.

3.2.1. Recurrent Congestion

The normal patterns of human and business activities create cyclical and recurring time periods when overall traffic on a network significantly increases. For example, parts of an ISP network with a high concentration of business users are likely to see higher usage during business hours than at other times. In contrast, parts of the network with a high concentration of residential users are likely to see higher usage during evening hours, as illustrated previously in Figure 4.

Recurrent congestion will typically last for multiple hours and when it occurs generally displays a periodic pattern, for example, weekday afternoons, every evening, or some other recurrence clearly linked to underlying human behaviors. As a result, this type of congestion tends to be predictable. Comparing Internet data networks to highways, an analogy for recurrent congestion is the average traffic patterns according to the time of day, including normal delays during rush hours.

3.2.2. Predictable Events

Specific predictable events can be the cause of network congestion by creating unusual Internet demand in addition to the existing demand, either as sources or destinations of traffic. Again comparing Internet data networks to highways, an analogy for congestion caused by predictable events is the incremental traffic caused by planned events such as road construction. A variety of examples can illustrate this type of congestion in the Internet:

- ***Mass in-person event***

  A mass in-person event occurs where many people gather in one physical place: sporting events, conventions, or political rallies, for example. These events are more likely to occur in populated areas. Wireless networks are most prone to these events, since they typically involve network users physically coming together. These events generally last for hours, although some may last only a few minutes. Some events are planned far in advance, and are thus predictable, but some occur with little warning.

- ***Mass on-line event: Users accessing an Internet site***

  A mass online event occurs when many people try to reach a particular Internet destination at the same time or try to consume the same streamed content from a single source: streamed sporting events, live news events, the release of popular software, online ticket sales of popular events, or shopping at popular websites on major holidays, for example. This type of event is the virtual version of a mass in-person event. While the traffic to or from any individual user may be relatively small, the concentration of traffic at the destination may be large and thereby cause congestion. These events generally last for hours to days. While many mass on-line events are predictable, the popularity of any given event can be unpredictable.

- ***Mass on-line event: User-to-user communication***

A mass distributed online event occurs when many users try to communicate directly with each other at the same time, for example during major holidays. This type of event can cause congestion over a wide geographic region. Generally these events last for hours and are predictable.

3.2.3. Unpredictable Events

Specific unpredictable events can also be the cause of network congestion. A highway analogy for congestion caused by unpredictable events is the incremental traffic caused by unplanned events such as accidents. A variety of examples can illustrate this type of congestion in the Internet:

- ***Changes in routing***

While changes in routing of traffic typically decrease congestion, unexpected changes in routing of high volumes of traffic can increase congestion, e.g. when a large content provider changes the ISP from whom it purchases Internet access. These events may cause congestion at the boundary between two ISPs' networks, as the increased flow may exceed capacity. Addressing such congestion may involve manual changes to routing (on a time scale of hours) or discussions about interconnection agreements between the two ISPs (on a time scale of days to months).

- ***Emergencies***

Unexpected life-threatening or property-damaging events – earthquakes, hurricanes, floods, tornadoes, or major automobile traffic incidents, for example – can cause large increases in network traffic. The traffic comes both from the direct response to the emergency and often from the desire of users not directly impacted by the emergency to seek information. Dramatic weather can also shift demand among networks, for example because of people working from home due to impassable roads. These events tend to be localized, although the size of the geographic area covered can vary greatly depending on the nature of the emergency. While many such events last for hours, some can persist for days or weeks.

- ***Network Accidents and Failures***

Congestion can occur because of a temporary loss of capacity in the network due to failures or accidents. As with any technology, links or routers can fail. Failures may be due to hardware, the result of software bugs, or secondary effects of an emergency that causes a loss of power or spikes in power that damage equipment. For example, network links can be severed due to earthquakes, high winds, tornadoes, floods, fallen trees, construction activities, and automobile accidents.

The loss of network links or routers results in a decrease in capacity. The remaining parts of the impaired network may not have sufficient capacity to accommodate the redirected traffic, and congestion may result. In addition, portions of the network may lose connectivity, which can create further congestion if sources retransmit packets that did not arrive at their destinations.

Congestion resulting from accidents and failures can last from seconds to hours or days.

- ***Attacks***

Denial of Service (DoS) attacks occur when large amounts of traffic are transmitted to a particular Internet destination in an attempt to deny access to legitimate service requests. These attacks are intended to exhaust the destination's resources such as bandwidth, CPU or memory of the servers and other service-enabling devices [BGPMON]. DoS attacks can result in congestion at the intended destination and in some cases within the network that provides the destination's Internet access. In addition, these attacks can often cause congestion in a geographic region well beyond the target of the attack, for example in the networks of ISPs along the routes to the targets. Some ill-behaved applications can also mimic or have the same effect as DoS attacks. Attacks can last minutes, hours, days or even weeks.

Examples of these attacks include: SQL Slammer, a worm that spread so quickly it caused service disruptions or denial of service in large portions of the Internet as routers became overloaded and routing sessions failed [Guardian]; the DoS attacks that occurred during an Estonian government protest in 2007 [Estonia]; those that resulted from an alleged dispute between CyberBunker and Spamhaus in 2013 [Kamphuis]; and alleged ongoing attacks on financial infrastructure in North America and Europe [Atias]. An example of an application having the same impact as a DoS attack is a peer-to-peer file sharing application that consumes critical home router resources to such an extent that it may interfere with other applications.

3.2.4. Random Congestion

In addition to the events discussed above, congestion can occur because a number of users sharing a portion of a network simultaneously have high demand for a very short period of time. This random congestion is simply part of the statistical nature of traffic on the network, as illustrated previously in Figure 2. Part of the reason for this is that many applications are designed to fully utilize available resources by increasing usage up until congestion occurs, whether in the operator's network, the home or the connection between the two. This type of congestion generally has a duration from milliseconds to tenths of a second.

3.3. Location of Congestion

Congestion can occur on any link or router within the Internet. The link or router in a network path where demand is highest relative to capacity is called the *bottleneck*. Although congestion will occur on any link or router where demand exceeds capacity, it is likely that, when congestion occurs, the bottlenecks will be in relatively lower bandwidth parts of the network (access networks, for example) that connect to higher capacity parts of the network (the core ISP networks and the networks of ASPs). This follows from network design which attempts to optimize performance and cost, as capacity in access networks is generally the most expensive part of the network. Locations of potential congestion in ISP networks are:

- ***Wireless broadband access links and routers***

  Wireless access links and routers (supporting both mobile and fixed wireless broadband services) are susceptible to each of the types of congestion discussed in the previous section, particularly recurrent congestion due to busy hour demand, mass in-person events, and emergencies. Because of limited wireless spectrum, relatively high cost and complexity of adding wireless capacity, network signaling requirements, variability in bandwidth availability due to device mobility, and environmental factors, wireless access links may be the bottleneck. Congestion at these locations affects users in the geographical region served by the congested wireless access link. In addition, because wireless devices may automatically attempt to connect to any nearby access point, failures of wireless links or routers can cause the remaining links and routers to become congested, thus affecting users in a wider geographical region.

- ***Wi-Fi wireless broadband access link and routers***

  Wi-Fi wireless broadband networks are a special case of wireless broadband networks. In addition to the types of congestion faced by all wireless broadband technologies, their use of unlicensed spectrum can cause temporary reductions in capacity due to interference from adjacent Wi-Fi networks or other devices or networks operating in Wi-Fi frequencies. Congestion in a Wi-Fi network may only affect users of that network, or it may also affect users in nearby Wi-Fi networks.

- ***Wireline access links and routers***

  Wireline access links and routers are susceptible to each of the types of congestion discussed in the previous section, except those types of congestion that are caused by mobility. In particular, congestion may occur due to busy hour demand, accidents and failures, attacks, and randomness. Congestion at these locations will affect users that share the congested wireline access link or router.

- ***Core network links and routers***

  Core network links and routers are susceptible to recurrent congestion, mass distributed online events, emergencies, and accidents. However, because they have relatively high capacity and are shared by many users, they are less susceptible than access links to attacks and to random congestion. Because traffic in the core is averaged over a greater number of users and is therefore less bursty, recurrent congestion is often minimized because it is easier to predict and plan for. When congestion occurs at these locations, however, it affects users in a wider geographical region than does congestion in access networks.

- ***Network interconnection routers***

  Network interconnection routers that connect one ISP to another ISP's network or to a large ASP's network are susceptible to recurrent congestion, mass on-line events, and attacks. Because of economies of scale, recurrent congestion is often minimized by sufficient investment in capacity. When congestion occurs at these locations, it will affect users in a wide geographical region.

There are also potential bottlenecks in users' home or office networks. These bottlenecks are not discussed in this report, as the focus here is ISP and ASP networks.

3.4. Indicators of Congestion

Network operators are continually collecting measurements on the links and routers in their networks. Due to the large volume of packets passing through a link or router, network operators will commonly aggregate measurements. For instance, rather than recording the instantaneous demand every millisecond, a network operator may calculate and record the percentile demand over a period called the *measurement interval*. The length of the measurement interval significantly affects the resulting indicator of congestion. Demand averaged over 10 second intervals will not show congestion whose duration is on time scales of milliseconds or tenths of a second, in this manner "smoothing" the resultant demand curve. Demand averaged over five-minute intervals will appear to be even smoother than demand averaged over 10-second intervals. Measurement intervals of several minutes are useful for examining congestion on time scales of minutes or longer. Common choices for measurement intervals are in the range from 5 to 15 minutes for capacity planning purposes, and many network operators examine the 95th percentile of demand over such measurement intervals.

One of the best predictors of congestion is the ratio of demand (averaged over a chosen measurement interval) to the capacity of a specific link or router, called *utilization* or *load*. When small measurement intervals are used, these measurements can be used to guide short-term congestion management. When longer measurement intervals are used, these measurements can be used to guide long-term congestion management.

Figure 5 shows an example of the percentage of link utilization on an access link interface during a one-hour time period using different measurement intervals. In addition to measuring demand, network operators commonly measure the average time from the arrival to the departure of a packet at a link or router (delay), and the proportion of packets that are not transmitted (packet loss).

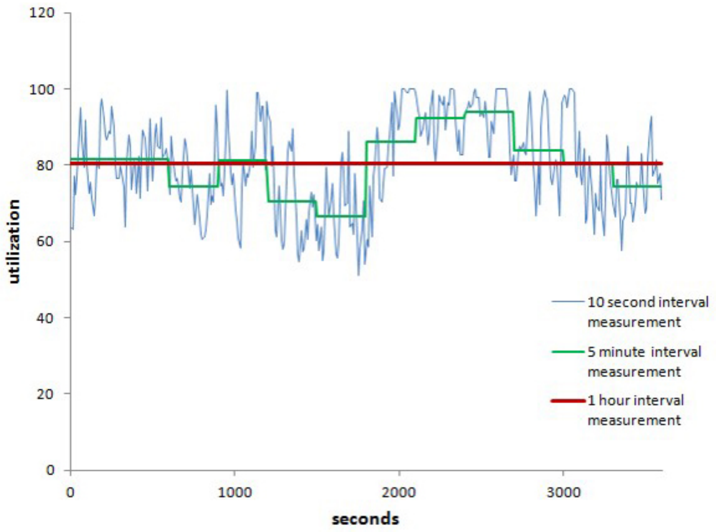

Figure 5. Link Utilization for Various Measurement Intervals

Congestion can also be measured by the effect that it has on a user's application. These are called *Quality of Service (QoS)* metrics. The most common are:

- *End-to-end packet delay*: The time from the transmission of a packet at the source to its reception at the destination. This includes delay due to (a) the time for a signal to propagate along the links, (b) the time for transmission of a packet at each router, and (c) queuing time due to congestion.
- *Delay jitter* (or simply "jitter"): The variation in end-to-end delay between packets.
- *End-to-end packet loss*: The proportion of packets that are transmitted by a source that do not arrive at the destination.
- *End-to-end throughput*: The average number of bits per second that are received at the destination.

3.5. Impact of Congestion on and by Applications

The satisfaction of a user with the performance of an application is called *Quality of Experience (QoE)*. When increased end-to-end packet delay, delay jitter, or end-to-end packet loss cause a degradation in QoE, it is generally noticed by the user in a variety of ways. Examples include:

- Increased response time of all or specific Internet applications.
- Webpages or parts of webpages (images, for example) take an increased time to load.
- Streaming audio or video suffers from decreased sound or picture quality, or is interrupted.

- Real-time audio (such as voice calls) or real-time video (such as video chat) suffers from decreased sound or picture quality or from unacceptable delays between speaking and hearing, or is interrupted.
- In multiplayer games, players may notice an increased delay between actions taken on the controller or home device and the results of these actions on the screen.
- Increased file transfer times.

Congestion may or may not be noticeable to users, depending on the application type, the application design, the severity of congestion as measured by QoS metrics, and the duration of congestion. The characteristics of an application and its design determine when and how QoE is degraded by congestion. Applications can be roughly classified by their sensitivity to QoS metrics and to the duration of congestion:

- ***Delay-intolerant applications:***

  Applications that are highly interactive are likely to have a QoE that is very sensitive to end-to-end delay. For example, the QoE of voice calls, video chat, and many multi-player games suffers when the end-to-end delay exceeds a few tenths of a second. Consequently, delay-intolerant applications will usually not request that dropped packets be retransmitted from the sender and will usually throw away packets that do not arrive within a certain time interval. Even brief occurrences of congestion, on the order of a few tenths of a second, can cause noticeable degradation in the QoE of these applications.

- ***Jitter-intolerant applications:***

  Applications that support synchronous communication are likely to have a QoE that is sensitive to delay jitter on the order of a few seconds or less. Since jitter is caused by variations in delay, all delay-intolerant applications are also jitter-intolerant. In addition, applications that stream audio and video are often jitter-intolerant. Consequently, jitter-intolerant applications will buffer received packets to equalize their end-to-end delay and thereby reduce their delay jitter. Real-time communications applications will only buffer packets for at most a few tenths of a second. In contrast, streaming applications often buffer packets for a few seconds. Time periods in which instantaneous demand exceeds capacity, that last longer than the buffering capability of a jitter-intolerant application, can cause noticeable degradation in its QoE.

- ***Minimum throughput applications:***

  Some applications are designed with the assumption that they will experience at least a certain minimum throughput. For example, the QoE of voice and video often degrades significantly if the throughput falls below the designed threshold. If the content is encoded in multiple formats, the application may respond to congestion by changing to a format with a lower throughput threshold, which reduces the QoE but is often preferable to not receiving the content at all. Minimum throughput applications can experience noticeable degradation in QoE when instantaneous demand exceeds capacity during time periods on the order of tens of seconds.

- ***Loss-intolerant applications:***

  Applications have different tolerances for end-to-end packet loss. Applications that expect data to arrive accurately and without error do not tolerate packet loss. Streamed video is an example of this, as users find missing pixels, frozen frames, and other manifestations of lost packets to provide a very unpleasant QoE. Other applications can tolerate a small amount of end-to-end packet loss, for example some audio and video applications. Applications that are loss-intolerant but delay jitter-tolerant, such as text messaging, text chat, email, and web browsing, will request that dropped packets be retransmitted to ensure that every packet is eventually received at the destination, even if some packets require multiple attempts. Degradation in the QoE of these applications will typically occur only if the duration of congestion is seconds to minutes. Congestion of any duration may delay file transfer times for bulk data transfer applications (such as software updates or peer-to-peer file sharing), with delays in proportion to the duration of congestion.

In summary, occurrences of congestion for a few tenths of a second may degrade delay-intolerant applications, occurrences of congestion for a few seconds may degrade jitter-intolerant applications, and occurrences of congestion for tens of seconds may degrade minimum-throughput applications. If the duration of congestion is on the order of minutes to hours, then it negatively impacts most applications. An operator's network design as well as congestion management practices affect whether and how congestion impacts various applications.

Network design can significantly affect the QoE of various applications. For example, the size of buffers in routers can affect the delay and loss of incoming traffic to each queue (see Section 5.7 on packet scheduling). The use of larger queues can decrease packet loss but increase delay and jitter.

The behavior of applications can either decrease or increase congestion. Some applications respond to congestion by decreasing their sending rates, thereby decreasing congestion. Some delay-tolerant applications may not decrease their sending rates but may use protocols that allow traffic transmissions to be scheduled in a manner that decreases congestion. There are other applications, however, that do not react to congestion or that react in a manner that increases traffic, e.g. by requesting retransmission of all delayed or dropped packets, without lowering the rate of transmission.

4. Classification of Congestion Management Techniques and Practices

This section of the report provides a classification of congestion management techniques and practices and delineates the scope of the report. Section 4.1 distinguishes between congestion management techniques and practices, or in other words the application of those techniques to effectuate a particular outcome. Section 4.2 classifies congestion management techniques by the time scale on which they operate. Section 4.3 briefly outlines the parties or organizations that may implement congestion management practices. Section 4.4 discusses when congestion management practices are based on user- or application- based criteria. Section 4.5 delineates the scope of the remainder of the report.

4.1. Techniques versus Practices

It will be helpful in this report to distinguish between congestion management techniques and specific implementations of those techniques, which are referred to as "practices", or in other words the application of those techniques to effectuate a particular outcome. This report uses the term *congestion management technique* to refer to a specific congestion management function that determines whether Internet traffic is transmitted or the rate at which traffic is transmitted, or that enables such functionality in other techniques. The techniques considered in this report include packet scheduling, packet dropping, routing, rate control, caching, resource reservation, and admission control, which are discussed in Section 5.

This report uses the term *congestion management practice* to refer to the use by a party or organization:

- of a collection of traffic management techniques,
- targeted at particular users and/or applications,
- upon the trigger of some event.

The parties or organizations that engage in or implement congestion management practices include Internet Service Providers (ISPs), Application Service Providers (ASPs), operating systems developers, customer premises equipment manufacturers, and consumers and enterprises. For example, an ISP may combine a set of congestion management techniques with the goal of reducing congestion for all users and applications, or with the goal of reducing congestion only for a subset of users or applications. In the latter case, congestion management may be used to differentiate products. An ASP that operates a network, for example, may combine a set of congestion management techniques with the goal of reducing congestion for all of its users and applications, or with the goal of reducing congestion only for a subset of its users or applications.

The Internet is based on the concept of a layered architecture, where each layer provides certain functionalities [RFC 1122]. A *layer* is an abstraction that hides the implementation details of a particular set of functionality. Congestion management can be applied in any of the layers. (A definition of each layer can be found in the Glossary.) Packet scheduling is commonly implemented in the lowest three layers: the *physical*, *link*, and *network* layers. Packet dropping and routing are commonly implemented in the network layer. Rate control and admission control are commonly implemented in the highest two layers: the *transport* and *application* layers. Caching is commonly implemented in the application layer. Resource reservation may be implemented at any layer, but is often controlled by decisions at the transport and application layers. Techniques that use deep packet inspection (DPI) are usually operating at multiple layers.

Section 5 of this report presents a survey of congestion management techniques to illustrate the range of techniques used by network operators. Section 6 presents a survey of congestion management practices of network operators.

4.2. Time Scales

Since congestion occurs on time scales from milliseconds to hours, as discussed in Section 3.2, congestion management techniques are also designed to work on a number of different time scales.

- ***Months to Years – Capacity Planning; Internet Subscription Plans***

  Capacity planning and augmentation occurs on a time scale of months to years. ISPs face rapidly growing demand for capacity. During the last year, average Internet demand per customer during the busiest hours in North America grew at an annual rate of 39% on fixed access lines and 25% on mobile access networks according to one estimate [Sandvine2013]. During the next five years, average Internet demand during the busiest hours in North America is expected to grow at an annual rate of 23% according to one estimate [CiscoVNI] and mobile Internet traffic at an annual rate of 40% according to another [Sandvine2013]. Network operators typically estimate demand months to years in advance, and use such demand estimates combined with the cost of capacity to plan a schedule for capacity upgrades. Since it may take months to implement a capacity upgrade, the time scale for such congestion management is months to years. Thus, although capacity planning can greatly affect how much congestion occurs on multiple time scales, it cannot react to congestion as it occurs.

  Internet subscription plans also affect congestion on a time scale of months to years. Internet subscription plans commonly include limits on downstream and upstream transmission rates. Since demand usually increases with transmission rates, the number of subscribers to a particular plan can affect demand and thus congestion. In addition, some plans include limits on the maximum number of bytes transmitted per month (commonly called *usage caps*) or charges for usage above some threshold. These limits can be viewed as long-term congestion management practices. Limits on transmission rates affect congestion in much the same way as do capacity decisions. Usage caps and usage charges may influence the amount of traffic users transmit over the course of a billing cycle. Although such limits can affect how often or the degree to which congestion occurs on multiple time scales, they cannot react to congestion as it occurs.

- ***Seconds to Minutes – Reservation; Prioritization; Rate Control; Routing***

Reservation, prioritization, and admission control techniques and practices can be used to react to congestion on a time scale of seconds to minutes, as well as to differentiate or guarantee performance on shorter time scales. Some networks allow resources such as bandwidth to be reserved or allow some packets to be prioritized over others. Prioritization techniques are usually implemented through packet scheduling in the link or network layers. Resource reservation may be implemented in any layer, but is often controlled by decisions at the transport or application layers. Typically the decisions that guide use of such practices are made on a time scale of seconds to minutes.

Rate control techniques such as those implemented in the Transmission Control Protocol (TCP) are designed to react to congestion on a time scale of a few tenths of a second or longer. Congestion management protocols residing at the transport or application layers are often used to limit the number of packets per second that a source application transmits. These congestion management techniques use information about end-to-end packet delays and losses, and thus dynamically update their limits on the time scale of an end-to-end delay (typically on the order of a few tenths of a second).

Although routing is typically based solely on the destination IP address, routes are sometimes adjusted in an attempt to minimize congestion. When these adjustments are made, typically it is on a time scale of minutes or longer. Packets generally progress through many routers before arriving at their destination. The path is determined by routers that exchange information concerning possible routes and congestion on these routes. In addition, sometimes content is available at multiple locations, and Content Delivery Networks can be used to balance the load and reduce congestion. This computation of routes is accomplished using network layer protocols. Routes may be updated as often as every few tenths of a second.

- ***Fractions of a Second – Packet Scheduling; Packet Dropping***

Packet scheduling techniques can be used to react to congestion on a very fast time scale. Packet scheduling techniques at the *data link layer* determine when to transmit each packet. When there is a queue of packets waiting for transmission, packet scheduling techniques also choose which packet to transmit next. This decision takes place each time a packet is transmitted, which can be roughly 100 to 1 million times per second, depending on the transmission rate and the packet size. The decision is guided by a simple algorithm that requires little computation. When a queue is full or nearly full, packets might be dropped. This decision takes places each time a packet arrives in a queue, which might also be roughly 100 to 1 million times per second.

4.3. Parties that May Engage in Congestion Management Practices

The parties or organizations that may engage in congestion management practices are:

- **Internet Service Providers**

  Each Internet Service Provider (ISP) implements a set of congestion management techniques at each router within its network. Thus, all applications communicating through the Internet indirectly rely on congestion management techniques implemented within networks of the ISPs through which traffic passes. Congestion management techniques within ISPs' networks includes routing (Section 5.5), packet dropping (Section 5.6), and packet scheduling (Section 5.7). Optionally, an ISP may also support admission control and resource reservation (Section 5.2). Certain networks, notably cellular data networks, also implement rate control (Section 5.4). Some ISPs also use deep packet inspection to classify traffic (Section 5.1).

- **Application Service Providers and Application Designers**

  Applications can, at times, implement congestion management techniques at each endpoint of the communication. At the user's endpoint, congestion management may be implemented within an application that a user runs on a device. At remote endpoints, congestion management may be similarly implemented in the communicating application at another location, within an application server, or within an ASP's own network. Congestion management is frequently implemented by applications that are moderately to highly interactive, including video streaming, voice over IP (VoIP), video conferencing, and gaming. Congestion management techniques within applications may include admission control and resource reservation (Section 5.2), caching (Section 5.3), and rate control (Section 5.4).

- **Operating Systems Developers**

  Each device's operating system (e.g. Windows, MacOS, Linux, Android, iOS) implements a set of congestion management techniques at each endpoint of the communication. Since all applications communicate by relying on the network functionality built into the operating system, all applications indirectly rely on congestion management techniques implemented within the operating system. Congestion management techniques within operating systems may include protocol support for allowing applications to request admission control and resource reservation (Section 5.2), application interfaces to TCP rate control capabilities (Section 5.4), and allowing applications to access and control protocol header information that can impact packet dropping (Section 5.6) and packet scheduling (Section 5.7).

- **Customer Premises Equipment Manufacturers**

  Customer premises equipment (CPE) consists of user end devices (including computers, smartphones, tablets, Internet-connected printers, Internet-connected digital video recorders, and Internet-connected game systems) and user networking devices (including cable modems, DSL modems, home routers, and home gateways). Each piece of CPE implements a set of congestion management techniques. Thus, all applications communicating from one device to another or to a server through the Internet indirectly rely on congestion management techniques implemented within CPE. Congestion management techniques implemented within CPE as part of the device driver includes packet scheduling (Section 5.7).

- **Consumers and Enterprises**

Both residential consumers and enterprises may implement congestion management techniques in their networks. For both consumers and enterprises, such techniques may be activated and configured within applications, operating systems, and CPE. Large enterprises may also implement congestion management techniques in a manner similar to ISPs.

4.4. Which Traffic is Subject to Congestion Management

One of the key design questions about any congestion management practice relates to the subset of network traffic with which the practice is concerned. Network operators target all traffic on their networks with some practices, whereas with other practices they target only the traffic of specific users, a subset of types of applications, all instances or specific instances of applications, or specific components of such applications.

User-based congestion management is applied to all traffic associated with a particular user or user group. Some ISPs or ASPs may define user groups based on:

- the service plan to which users are subscribed (e.g., all users subscribed to an ISP's basic broadband Internet access plan, a cellular provider's unlimited data plan, or an ASP's premium product);
- the volume of data that users send or receive over a specified period of time or under specific network conditions (e.g., all users who consume 300 GB in a month, the top 5% of users by data consumption during a busy period on the network, or users who are in the process of transmitting the first 20MB of a file); or
- the location of users (e.g., all users in a particular geographic area experiencing an emergency).

User-based congestion management does not require network operators to examine the content of network traffic as the decision to apply user-based management is agnostic to that content – it depends only on which users are generating traffic, not what they are generating.

Application-based congestion management is applied to all traffic associated with particular uses of the network. That is, congestion management is application-based if network operators select traffic to be managed because it:

- has a particular source or destination (e.g., http://www.example.com);
- is generated by a particular application (e.g., a BitTorrent client);
- is generated by an application that belongs to a particular class of applications (e.g., video chat applications that include Skype, Google Talk, WebEx, and FaceTime);
- uses a particular application- or transport- layer protocol (e.g., Session Initiation Protocol, User Datagram Protocol, or Hypertext Transfer Protocol); or
- is classified for special treatment by the user, application, or application provider (e.g. traffic identified by the user's application as delay-intolerant, or traffic identified by the application provider as jitter-intolerant).

Application-based congestion management depends on the network operator's ability to identify the traffic associated with particular uses of the network. Techniques used by network operators to identify and select traffic subject to an application-based congestion management practice might be based on packet payloads (using deep packet inspection or other content-aware network equipment), network or transport layer headers (e.g., port numbers or priority markings), heuristics (the size, sequencing, and/or timing of packets), or a combination of these characteristics.

Congestion management may also be both user- and application- based. For example, a network operator could choose to rate-limit video streaming for all users who consume 300 GB in a month or for all users in a congested cell in a cellular network.

User- and application- based congestion management may be based in part on economic and legal agreements between network operators or between users and network operators. *Service Level Agreements* (SLAs) between network operators delineate contractual aspects of the service, often including the upstream and downstream bit rates at the boundary between the operators' networks, the maximum delay across an operator's network, sometimes the maximum proportion of packets that may be dropped or other QoS metrics, and sometimes specifications of payments.

In contrast to both user-based and application-based management, some practices apply to all traffic regardless of user group or application. For example, a network operator might program its routers with dynamically adjusting buffers to accommodate rapid changes in network load (Section 5.6). This practice applies to all traffic on the network and can help to mitigate the impact of congestion regardless of which users' traffic is flowing through the routers or which applications are in use. Similarly, network operators might make decisions about traffic scheduling, queuing, or routing based on factors unrelated to users or applications – how quickly packets have arrived at a particular point in the network, or which ones arrived first, for example. These kinds of practices are user- and application-agnostic.

4.5. Scope of the Remainder of the Report

The remainder of this report focuses on user- or application- based real-time network management of Internet services. This is an important set of congestion management techniques and practices. It is also a set that has generated public policy discussions. In order to be considered in the remainder of this report, a congestion management technique or practice must:

(1) Be implemented or potentially implemented by a network operator. As discussed in Section 4.3, other parties or organizations that may implement congestion management practices – such as applications, operating systems, and device drivers – are important; however they are outside the scope of this report.

(2) Act on a time scale of minutes or less. As discussed in Section 4.2, congestion management techniques and practices that operate on a time scale greater than minutes (e.g. capacity upgrades, limits on downstream and upstream transmission rates, usage caps and usage charges) may influence the amount of traffic users transmit over the course of a billing cycle; however they are outside the scope of this report.

(3) Be used for purposes of congestion management. Similar techniques may also be used for other purposes, including security; however use for such non-congestion management purposes is outside the scope of this report.

(4) Be user- or application- based. Congestion management techniques and practices whose goal is to reduce congestion for a chosen set of users or applications will be considered. As discussed in Section 4.4, there are many congestion management techniques whose goal is to reduce congestion for all users and all applications; however they are outside the scope of this report.

This report focuses on congestion management techniques and practices that pass all four of these tests. However, the BITAG notes that congestion management techniques and practices that do not pass all four of these tests are important, and it may consider them in future reports.

5. Congestion Management Techniques

As noted in Section 4.1, this report uses the term *congestion management technique* to refer to a specific congestion management function that determines whether Internet traffic is transmitted or the rate at which traffic is transmitted, or that enables such functionality in other techniques. Many congestion management techniques at or above the network layer are standardized by the Internet Engineering Task Force (IETF), and many at the physical and data link layers are standardized by the Institute of Electrical and Electronics Engineers (IEEE) or by industry consortia (e.g. CableLabs, 3GPP). Standardization can allow for interoperability and for consistent functionality in network devices and equipment. Other congestion management techniques have not been standardized or are propriety.

This section focuses on congestion management techniques and for each technique explains who may apply the technique, the duration and location of congestion the technique addresses, and the intended impact upon applications. Section 5.1 illustrates how either a user or a network operator may classify a packet. Section 5.2 discusses congestion management techniques designed to support applications that require a minimum amount of network resources in order to function at the desired performance level. Section 5.3 discusses congestion management techniques that temporarily store (or *cache*) popular content in multiple locations. Section 5.4 discusses congestion management techniques that control the average rate at which a source transmits traffic into the Internet. Section 5.5 considers routing and traffic engineering. Section 5.6 discusses how a router may decide when to drop a packet and/or *mark* it to indicate congestion. Section 5.7 examines how a router places packets into queues and the order in which it transmits packets. Finally, Section 5.8 discusses how these techniques may be combined to offer a collection of capabilities in various QoS architectures and in various access network architectures.

Each of these sections focuses on congestion management techniques that may be user- or application- based. Congestion management techniques that are agnostic to both user and application are very commonly implemented but are not generally discussed here. Section 6 illustrates congestion management practices that use many of these techniques.

5.1. Packet Classification

In order for a congestion management technique to be based on a particular user or a particular application type, application, or application component, packets must be classified using criteria that identifies the particular user or application traffic. This section illustrates how users, ASPs, and ISPs can classify packets.

Most packets in the Internet are transmitted by network operators using *best-effort service*, in which routers transmit packets in the order in which they are received, and without regard to the source, the application that generated them, or the resulting QoE (as discussed in Section 3.5). Hence best-effort service requires only knowledge of the destination of a packet, not of the source, the application, or the desired QoS characteristics.

In contrast, user- or application- based congestion management may require knowledge not only of the destination but also of the source, user behavior, the type of application, application, application component, the user's or application's desired QoS characteristics or QoE, agreements between network operators, or agreements between users and network operators (as discussed in Section 4.4). Each user, ASP, and ISP may classify packets on the basis of such information. A *classifier* is an entity that selects packets based on the content of packet headers or other attributes according to defined rules [RFC2475]. Classification of packets may be performed using attributes from any of the following layers: application, transport, network, or data link. This report uses the term *flow* to indicate a group of packets that share a common set of properties [RFC5472]. One purpose of classification is to allow a network operator to apply congestion management techniques based in part on a packet's classification.

A packet's classification is often used to decide how to apply congestion management techniques discussed in the remainder of this section. Classification permits the network operator to make choices, for example ensuring that applications receive the desired treatment, to force "aggressive" applications (e.g., applications that attempt to use all available resources) to make way for less aggressive applications, or to ensure that bandwidth contracted for in an SLA is available for the intended purpose.

Markings that indicate the classification of a packet are often placed in specific bits of a packet header to allow simplified subsequent classification based only on these specific bits. The classification markings can allow a user, ASP, or ISP to indicate a desired packet treatment. For example, a source can use a classification marking to request that the packet be treated in a manner consistent with expectations for Voice-over-IP traffic [RFC 2474, 2475, 4594], as discussed below in Section 5.8.2. Alternatively, the classification marking can be used to indicate packets to or from a specific host or network, or traffic related to a given application, as discussed below in Sections 5.2 and 5.5. Classification markings can also be used to uniquely identify a flow, so that priority can be given on the basis of the throughput experienced by a flow.

A classification marking can be placed in the packet header corresponding to the protocol that generated it. The classification marking may indicate a unique identifier of the flow. For example, the Multiprotocol Label Switching (MPLS) protocol (discussed below in Section 5.5) includes fields in its packet header that can be used to classify packets and to assign the route (or portion of a route) taken by those packets. Alternatively, the classification marking may indicate the desired treatment of the packet. For example, the DiffServ architecture (discussed below in Section 5.8.2) includes a *codepoint* in the IP packet header that can be used to classify packets of *differentiated services* (e.g., network control, VoIP, video streaming, best effort traffic) in order to provide desired routing, scheduling, and dropping treatment. Data link layer protocols often include similar classifications, e.g. the Ethernet protocol includes a field in its packet header that can be used to classify packets for similar purposes [802.1Q][802.1p].

Each user, ASP, and ISP may classify, mark, and re-mark packets according to its architecture, objectives, and agreements. Thus a packet's classification marking may be modified as it progresses from source to destination. Most commonly, marking and re-marking may occur at user-operator and operator-operator boundaries, according to the agreement or lack thereof between the parties. For instance, a network operator may classify a packet upon ingress to its network by inspecting the MPLS fields, the DiffServ codepoint, the Ethernet fields, or another data link layer codepoint. It may also classify packets based on the transport protocol, source IP address, source port, destination IP address, or destination port. If a network operator does not support packet classification or does not have an agreement with a user, ASP, or ISP to honor a packet's classification marking, it may re-mark the packet to receive best-effort service; alternatively a packet's classification marking may be preserved but ignored by the network operator.

One challenge to packet classification is the ease or difficulty in accurately identifying particular application traffic. The applications themselves may know their QoE requirements. However, packet classification may face additional challenges, since some application traffic may not be easily identified. For example, an ISP may have difficulty differentiating between a normal VoIP call, an emergency VoIP call, and peer-to-peer traffic, that all use the same set of protocols. Deep Packet Inspection (DPI) is a method that is sometimes used to better identify the application and its characteristics by looking for certain patterns inside a packet or amongst packets in a flow. For purpose of congestion management, this technique is not useful on encrypted content, and can be of limited use where application developers purposefully attempt to thwart DPI by removing or modifying these specific characteristics and patterns.

In general, classification based on attributes other than protocol elements intended for classification marking (e.g., MPLS fields, DiffServ codepoint, Ethernet fields, or IPv6 flow label [RFC 6437]) is only done at the edge of an operator's network. Internal network routers then rely on the marked classification. However, reclassification and re-marking may also occur within an operator's network based on congestion encountered at routers, or because packets associated with reserved resources are exceeding the allocated resource (see Section 5.2).

5.2. Admission Control & Resource Reservation

As discussed in Section 3.5, some applications are delay-intolerant, jitter-intolerant, or require a minimum throughput, e.g. voice over IP communications, high definition video conferencing, and video streaming. Depending on the nature and design of the application, congestion on the order of a few tenths of a second to tens of seconds can cause noticeable degradation in QoE for such applications.

These applications by design require a minimum level of network resources, e.g. bandwidth, in order to function adequately. Resource reservation is thus a congestion management technique appropriate for such applications. Reservation of an appropriate level of resources can ensure that congestion does not negatively impact the flows using the reserved resources. However, flows that are sharing the remaining unreserved resources will suffer a consequently larger impact of congestion upon their performance than if no resources had been reserved and if these resources had been made available to all traffic.

The offering and granting of resource reservations is a decision made by each network operator. Resource reservations require that routers be capable of responding to application requests for special treatment of certain flows, e.g. minimum bandwidth or maximum delay or delay jitter. The challenge for a network operator that chooses to use resource reservations is to allocate resources in a manner that optimizes their use.

Resources may be reserved (a) only in portions of a single operator's network in which congestion more frequently occurs, (b) throughout a single operator's network, or (c) cooperatively amongst multiple network operators through business arrangements. Although business arrangements between enterprises and ISPs regarding resource reservations are not uncommon, business arrangements between ASPs and ISPs or between multiple ISPs regarding resource reservations are currently rare.

If resource reservations are offered, a network operator must also have the capability to control which flows are granted reservations (called *admission control*). A network operator may choose to accept a request, deny a request, or issue a counteroffer for a reduced quantity of resources. These decisions are automated, and made on a very short time scale.

A number of standards have been developed to support resource reservation and admission control, see e.g. IETF's Internet Services (described further in Section 5.8.1), 3GPP's Policy and Charging Control (described further in Section 5.8.5), CableLabs' PacketCable DOCSIS QoS (described further in Section 5.8.3), Frame Relay, and Asynchronous Transfer Mode (ATM).

It is also possible to statically reserve resources for certain users or application traffic by configuring the capacity allocated to that traffic. This configuration can be done by allocating capacity (such as frequency spectrum) on a physical medium, by allocating link capacity (such as an ATM Permanent Virtual Circuit [PVC]), or by allocating capacity through rate control (see Section 5.4) or through packet scheduling (see Section 5.7).

5.3. Caching

As discussed in Section 3.3, core network links and routers are susceptible to recurrent congestion and mass distributed online events. As discussed in Section 3.5, such congestion can increase the download time of content. One technique to address this issue is to cache (or store) popular content in multiple network locations. Caches may be placed in several types of locations, including the ASP's network or attached to or in multiple ISPs' networks.

Caching is typically implemented by directing a request from a user for a particular piece of content to the closest cache that has the content or to the most lightly loaded cache. If the closest cache is selected, this lessens traffic between the ASP and the user, including core network links and routers. When caching is combined with network measurement systems to form Content Delivery Networks (discussed in Section 5.8.6), the resulting congestion management practice can lessen traffic on congestion network links and routers.

The direction of a request to a cache is typically dependent upon the particular content requested. Since the content is related to the application, one may consider this an application-based technique.

5.4. Rate Control and Traffic Shaping

As discussed in Section 3.1, congestion occurs when the sum of users' individual instantaneous demands exceeds the network capacity of a link or router. Many applications use the Transmission Control Protocol (TCP), which controls the rate at which packets are transmitted in an attempt to limit congestion. This type of *rate control* is handled directly by the TCP implementations at the application endpoints, but can be influenced (in a user- and application- agnostic manner) by packet dropping and scheduling techniques implemented in provider networks.

However, sometimes it is useful to apply rate control on the basis of the user or application criteria. In these cases, it may be applied by the applications directly, operating systems, ASPs, or ISPs.

Rate control by applications and operating systems is implemented at the source and destination. It may be based on the requirements of the application. For example, many streaming video applications do not use TCP in order to avoid TCP's rate control, since the rate control may unacceptably increase the delay jitter. Instead, they often choose to use the User Datagram Protocol (UDP), which does not adapt transmission rates based on congestion. Thus, a streaming video application may implement its own rate control techniques to adapt its transmission rate based on network congestion. The IETF has developed a pair of protocols, Real-time Transport Protocol (RTP) and the RTP Control Protocol (RTCP) [RFC3550], that can be used by applications to control the transmission rate and QoS characteristics of real-time applications such as VoIP and streaming video.

ASPs often use similar rate control techniques in their servers to ensure the QoS characteristics of real-time applications. In addition, an ASP may use rate control on the basis of user to ensure fair sharing of capacity.

Network operators may use similar techniques called *traffic shaping*. Traffic shaping by network operators is implemented at routers between the source and destination, most commonly at the edge of an operator's network. For example, a network operator may use traffic shaping on the basis of user to limit congestion caused by heavy users or to enforce the service tier rate. Traffic shaping is accomplished by intentionally delaying the transmission of selected packets. Traffic shaping is often used in conjunction with packet classification to limit the capacity used by traffic placed in various classifications. A common method involves placing packets from a selected flow into a queue, and serving that queue at a fixed rate, thereby limiting that flow to a maximum capacity. Alternatively, but more rarely, when a flow exceeds this maximum capacity, packets may be dropped or reclassified to indicate that they are not within the allowed limit (a technique known as *policing*). A network operator using traffic shaping determines which flows are subject to traffic shaping and the corresponding capacity limits as part of a traffic management practice, as discussed in Section 6. This technique is often applied to traffic that has been provided reserved resources (see Section 5.2) when that traffic exceeds the reserved limit.

A less common form of rate control by ISPs is *transcoding*, in which content compressed at the source is decoded and re-encoded by a router into a flow at a lower rate. This technique may be applied by an ISP when a flow is progressing from a network with a higher capacity to a network with a lower capacity.

5.5. Routing and Traffic Engineering

Each router in the Internet participates in determining the *routes*, or sequences of routers and links, that a packet traverses from each source to each destination. Within an operator's network, a standardized protocol such as Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS) is typically used to efficiently get each packet to its destination (if within the operator's network) or to the next operator's network. The selection of a route within an operator's network is typically determined solely on the basis of the destination IP address, but routes may be adjusted in an attempt to minimize congestion. The choice of route is thus agnostic to both user and application.

Inter-provider routes – those that traverse multiple operators' networks – are additionally determined using a standardized protocol, Border Gateway Protocol (BGP), that allows each network operator to implement its own SLAs and manage the routing that it will accept. As with OSPF, BGP usually determines routes solely on the basis of the destination IP address, and is usually agnostic to both user and application. Such routing seeks the most cost-efficient route from each individual provider's perspective, or a route that meets a particular contractual requirement. These contracts often seek to minimize congestion on inter-provider links.

This approach works well for many customers. However, in order to fulfill more comprehensive or detailed SLAs, the network operator must know the amount of data it will be expected to support and the behavior of subsequent paths. Queuing and queue management algorithms (such as discussed in Sections 5.6-5.7) can manage short-term congestion. However, if there are multiple potential routes to a destination and if the network is heavily loaded, OSPF and BGP may not effectively balance the traffic among those routes. Although adding capacity may solve this issue, it is often more economical for a network operator to exercise greater control in determining traffic routing and therefore the use of available capacity.

An alternative approach to determining routes solely on the basis of the destination IP address is to route certain traffic to the destination based on the type of contract or the desired QoS characteristics. For example, a network operator might sell several different classes of contracts, and determine the route based on the combination of destination IP address and contract class. One set of techniques for accomplishing this is called policy-based routing; a standard implementing this is Multi-Topology Routing [RFC4915]. Another common set of techniques, generically called virtual circuit routing, allows a network operator to determine labels on the basis of its SLAs; it has been standardized using a protocol called Multiprotocol Label Switching (MPLS). A network operator using MPLS might assign virtual circuits from each network ingress point (or each customer at an ingress point, or each service at an ingress or point of origination) to each network egress point so that the expected traffic load on each virtual circuit does not exceed a target threshold [RFC3272]. This approach, called traffic engineering, is used in most large transit networks. In addition to economically managing recurrent congestion, traffic engineering can also be used to react to network accidents and failures if a network operator sets up backup virtual circuits using disjoint paths to the primary virtual circuits, and moves traffic from the primary virtual circuit to the backup virtual circuit if there is a failure on the primary route. There are also similar traffic engineering techniques at the data link layer in some access networks, e.g. Virtual LANs in Ethernet architectures [802.1Q].

Another common use of traffic engineering is to satisfy SLAs for particular customers and services; a network operator might use one set of virtual circuits for general Internet traffic and another set of virtual circuits for each of a number of customers' internal networks, all using the same physical network. MPLS can also be combined with IntServ (discussed in Section 5.8.1) to create a scalable traffic management architecture that supports differentiated services [RFC2430]. In these cases, routing may be user- or application-based.

Traffic engineering for routes traversing multiple operators' networks can also be accomplished using a technique called BGP Traffic Engineering. The BGP protocol allows a network operator to determine which routes through its network it offers to customers and to interconnecting providers. Using BGP Traffic Engineering, a network operator may more tightly control which routes are used by which customers and interconnecting providers. This may allow the network operator to reduce the portion of an end-to-end route that traverses its network, thereby transferring costs to other providers.

Application Service Providers (ASPs) may also have a say in the route that the data will take to get to the desired destination, as well as the underlying quality of delivery. Large ASPs deliver their services using a distributed platform made up of clusters of servers that are hosted at one or more locations across the Internet. Large ASPs will often use some form of load balancing to distribute the requests for content across the clusters and the underlying network connections to the platform. The algorithm for determining where to map a request for content and its delivery is typically a function of the available processing capacity at a cluster, the current network conditions (i.e. speed, latency, throughput, packet loss, etc.) between the cluster and destination, and transit cost to reach the destination. This will often, but not always, result in an ASP mapping the content delivery request to a cluster with available capacity that will optimize both the delivery performance and cost for the ASP.

5.6. Packet Dropping

Routers receive packets on multiple incoming interfaces. Since the Internet uses statistical multiplexing, as discussed in Section 3.1, this incoming traffic usually encounters a queue of packets waiting to be transmitted on the same outgoing interface of the router, resulting in short delays. Each queue has a buffer of a certain size in which to store packets, and thus occasionally an arriving packet will find that the buffer is full. Routers thus must occasionally drop packets. Higher layers can retransmit dropped packets or notify applications that packets have been lost, as desired by the application. These higher layers can also restore packets into their original sequence if they have arrived out of order.

The simplest dropping technique only drops a packet when there is no space for it in the buffer. Packets may also be dropped based on the configured maximum data rate. Another tool available is to drop or mark packets when a buffer is nearly but not yet full in order to send a signal to the sources that they should reduce their traffic [RFC2309][RFC3168][RFC6679]. If done properly, this last technique can achieve a good balance between delay and throughput. For example, the Active Queue Management dropping techniques can be used to maximize average throughput while minimizing average latency during transient congestion events without unnecessarily targeting individual data flows or applications [RFC2309].

The simplest queuing technique places arriving packets in queues in the order in which they arrive. The simplest dropping technique drops the newest packet arrival when a drop is determined to be necessary or to send a signal. Under this system, packet drops will thus be agnostic to user and application. However, more complex dropping policies can drop packets based on classification, and by extension on user- or application- based criteria. For example, packets classified with a DiffServ codepoint intended for Voice-over-IP may be placed in a separate queue with a lower maximum queue length, and thus may experience a different dropping probability than best-effort packets.

Finally, a dropping technique may be used to discriminate against attack or illegal traffic. As noted in Section 3.2.3, networks occasionally experience a large influx of traffic as a result of an attack or other unplanned event (such as unforeseen popularity or an emergency event). Attacks and unplanned events can result in an arrival rate that significantly exceeds the router capacity. In this case, a network operator may choose to discard all traffic associated with an attack, or to limit traffic associated with an unplanned event, if it can accurately classify that traffic. Unfortunately, accurate classification of attack traffic may be difficult. It is also possible for a network operator experiencing a large influx of traffic to request other network operators to implement temporary policies in their routers to block or restrict traffic destined for the network experiencing the heavy traffic. If traffic is undesired, it is preferable to drop it as close to the source as possible to limit the impact it may have on all intermediate points.

5.7. Packet Scheduling

Each router implements an algorithm that determines how packets are placed into queues and the order in which the router transmits packets. Packet scheduling occurs primarily at the network and data link layers.

The simplest queuing technique places arriving packets in queues in the order in which they arrive, and the simplest packet scheduling technique transmits the packet at the head of the queue; this technique is called First-In First-Out, and it transmits packets in the order in which they were received, independent of user or application.

However, other packet queuing and scheduling algorithms can be used to ensure some type of fairness between users or to differentiate between classes of service. In this case, the packet queuing algorithm maintains multiple queues and places each packet into a queue based on its classification. The packet scheduler transmits packets from the queues to achieve the desired goal. There are many different algorithms in common use for selecting the next packet to transmit, e.g. Weighted Round-Robin and Weighted Fair Queuing. By varying the allocation of transmission capacity between the multiple queues, a router may differentiate the delay experienced by packets with different classifications. For example, packets with a DiffServ codepoint classification intended for Voice-over-IP may be placed in a separate queue from best-effort packets; if this queue obtains a proportionally higher capacity than that for best-effort packets, then the Voice-over-IP packets will experience a lower average delay than will best-effort packets.

Data link layer protocols include not only the types of packet queuing and scheduling algorithms described here, but also algorithms that determine when users may access a shared medium (e.g. a shared transmission line running down a street or a shared wireless channel). The algorithm is matched to the transmission characteristics of the physical network and the link layer protocol (e.g. Ethernet, DOSCIS, or a specific cellular link layer). As with packet scheduling, this decision can be made either on user- or application- based criteria, or agnostic to both. Commonly, the maximum throughput of a device on a link will be independent of user or application. However, many link layer protocols can be configured to give specifically identified traffic different shares of the link capacity, as discussed in Section 5.2.

If packet scheduling is implemented in a manner that is not agnostic to user and application, then it can be combined with packet dropping policies and admission control policies to provide different classes of service, as discussed in Sections 5.8.1-5.8.2.

5.8. Collections of Congestion Management Techniques

This subsection discusses how these techniques may be combined to offer a collection of capabilities in various QoS architectures and in various access network architectures. Sections 5.8.1 and 5.8.2 discuss QoS architectures that can be used in any IP network to create services with differentiated performance either within an operator's network or end-to-end. Sections 5.8.3-5.8.5 give an overview of access network architecture-specific implementations of congestion management techniques for broadband Internet access over cable, telephone, and cellular networks respectively. Section 5.8.6 gives a similar overview for Content Delivery Networks. In Section 6, congestion management practices that use many of these collections of techniques will be illustrated.

5.8.1. IntServ over IP networks

Resource reservation and admission control techniques were discussed in Section 5.2. The IETF has developed a QoS architecture called Integrated Services (“IntServ”) to support end-to-end resource reservation requests [RFC1633][RFC2205]. This architecture supports the ability of user applications to signal the operator’s network with requests for Quality of Service treatment. The design is based on a flow-based model wherein all or a portion of an application’s traffic (as determined by the application) is treated as a unique flow of data between two points. IntServ combines an end-to-end signaling protocol along with classification, resource reservation and admission control, routing, packet scheduling and dropping techniques. Routers between multiple operator networks may also communicate with each other to make decisions about resource reservation requests. It may be combined with routing techniques (e.g. MPLS, discussed in Section 5.5) to simultaneously reserve a network path and resources on that path.

5.8.2. DiffServ over IP networks

The IETF has also developed another QoS architecture called Differentiated Services (“DiffServ”) that uses packet classification, rate control, packet dropping, and packet scheduling techniques to implement scalable service differentiation in the Internet [RFC2474][RFC2475].

DiffServ includes a field in the header of IP packets that can be used to mark packets as belonging to a particular class. This allows applications and network operators to classify traffic into categories of their choosing. Network operators and applications can then use these DiffServ markings to differentiate among packets as they move through the network, giving some classes of traffic higher priority access to network resources than others.

Many network operators use DiffServ to prioritize the operator’s own network control traffic, to implement business agreements with enterprises that include requirements to achieve certain values of QoS metrics, or for their own VoIP and video streaming services.

DiffServ classifications and markings established by one network operator, however, will not necessarily be honored or retained when packets are handed off to another network operator. For DiffServ to work across multiple connected networks or on an end-to-end basis, special agreements are required between operators of interconnected networks. For example, two network operators may agree to configure the routers at the edge of their networks to classify and provide differentiated treatment to packets they receive from one another with selected DiffServ markings, so long as such traffic conforms to agreed-upon parameters regarding classifications and bandwidth. Currently, it is common that a network operator will reclassify a packet as best effort if there is no agreement that enables differentiated service.

DiffServ achieves scalability by having more complex classification done only at the edge routers of a network, with internal routers applying consistent dropping and scheduling behaviors. The edge routers determine how different packets should be classified, as well as whether and how to conduct admission control, rate control, or changing some packets' Diffserv marking to limit the volume of traffic assigned to each category. If too much network traffic is classified for priority treatment, prioritization may cease to assure the intended performance.

The forwarding treatment that DiffServ-compliant internal routers apply to packets with particular DiffServ markings (known as DiffServ codepoints, or DSCPs) are called per-hop behaviors (PHBs). PHBs are the building blocks of differentiated services; they are the mechanisms by which different traffic classes are granted different relative priorities in accessing network resources. If a group of PHBs is intended for general use, then it can be standardized. The DiffServ architecture has defined two initial PHB groups. The Expedited Forwarding PHB [RFC3246] is intended to provide a building block for low delay, low jitter, and low loss services that may support delay-intolerant and jitter-intolerant applications such as VoIP. The Assured Forwarding PHB Group [RFC2597] is intended to provide a building block to create services that are differentiated in terms of packet loss and delay, e.g., for jitter- and loss-intolerant applications such as video streaming. Additional per-hop behaviors may be defined in the future.

The end result of DiffServ is to allow packets belonging to a certain classification to experience different delay and/or packet loss than packets of a different classification.

While the use of DSCPs in conjunction with per-hop behaviors was defined as part of a cohesive QoS architecture, these components are often used outside of the cohesive architecture to classify traffic and to trigger particular scheduling or dropping behaviors.

5.8.3. Broadband Internet Access over Cable Networks

CableLabs, an industry consortium, has developed a network architecture called PacketCable [PacketCable] to implement Quality of Service enhanced communications over DOCSIS (the link layer protocol used in cable networks). PacketCable is a scalable QoS architecture that uses classification, admission control and resource reservation, packet dropping, and packet scheduling techniques, as discussed in Section 5. PacketCable leverages the QoS treatment capabilities within the DOCSIS protocol to enable packet classification and packet scheduling of IP flows that traverse the cable access network connecting a user's cable modem to the cable access router. Cable networks using the DOCSIS protocol utilize a pair of packet schedulers, one for the upstream (from the user to the Internet) and another for the downstream (from the Internet to the user), to schedule packet transmissions.

Routers in the upstream path in a cable network must statistically multiplex traffic from multiple users. They use a combination of packet classification, packet scheduling, and packet prioritization techniques to manage the rate at which packets are transmitted, and the order in which they are transmitted. An upstream controller takes turns between users who wish to transmit traffic. It can be used not only to support best-effort service, but also to reserve bandwidth for selected flows. The cable modem can classify packets based upon the IP header to determine a packet's scheduling and priority.

Routers in the downstream path in a cable network have a simpler task, in that all incoming flows arrive in a single queue. Correspondingly, they use a much simpler packet scheduler. It can also be used not only to support best-effort service, but also to reserve bandwidth for selected flows by classifying packets based on the IP header.

The upstream and downstream schedulers work together to allow service providers to define and implement a variety of IP-based services, including residential high speed Internet service, digital voice telephony, and service provider video-on-demand.

PacketCable enables an ISP to manage IP flows within the cable access network. If both the source and destination reside in a single cable access network, then this may be sufficient to guarantee QoS metrics. If, however, either the source or destination resides outside the cable access network, then PacketCable may be combined with other IP QoS architectures, e.g. IntServ or DiffServ, to guarantee end-to-end QoS metrics.

5.8.4. Broadband Internet Access over Telephone Networks

Digital Subscriber Line (DSL) and Passive Optical Network (PON) technologies used to offer broadband Internet access over telephone networks assign frequency spectrum on the physical medium to upstream, downstream, and network control traffic. This assignment defines the capacity available to each type of traffic. There is no other traffic differentiation implemented at the physical layer. This is quite different from the design approach of cable and cellular network technologies.

Differentiation of services can be achieved through various link and network layer mechanisms, such as the use of ATM PVCs, Ethernet VLAN tags, or other link and network layer classification techniques (see Section 5.1) that lead to differentiated packet treatment. Shaping techniques (Section 5.4) can be used to further define the actual bandwidth allocated to various traffic flows.

5.8.5. Broadband Internet Access over Cellular Networks

3GPP, a cellular industry consortium, has developed a network architecture called Policy and Charging Control (PCC) [3GPP PCC specification]. It uses packet classification, resource reservation and admission control, rate control, packet dropping, and packet scheduling techniques to provide QoS treatment to various services over cellular networks. A major difference between 3GPP architectures and some other access architectures is that they are connection-oriented as opposed to connectionless, which means that the manner in which the network responds to the QoS parameters is fundamentally different, even though the impact on an application's performance may be similar.

The objectives for providing QoS treatment in cellular networks are similar to those in other network architectures, but with additional challenges posed by the particular susceptibility of wireless broadband access technologies to congestion (as described in Section 3.3) and by packet loss due to environmental factors. 3GPP has defined a variety of mechanisms for providing QoS treatment, in both the radio access network and the associated core network. The need for resources may be indicated and the resources reserved either by user equipment or the wireless ISP's network elements, using 3GPP-defined parameters. These parameters provide the ability to request a desired QoS treatment (e.g. packet loss or delay), to control packet forwarding treatment (e.g., scheduling weights, admission thresholds, and queue management thresholds), and to control resources (e.g. using access control, prioritization, and preemption). The 3GPP PCC standard allows an ISP to charge for QoS treatment and usage, and to classify packets on the basis of user or application criteria or the terms of a subscriber's contract.

In addition to the 3GPP mechanisms, ASPs interfacing with smartphone apps have specialized methods for optimizing wireless resources. These include the use of push notifications versus polling, transmission scheduling and bundling, transcoding for content optimization to the user equipment, and local caching in the user equipment [GSMA].

5.8.6. Content Delivery Networks

Content Delivery Networks (CDNs) provide services for ASPs that can be used to improve the QoE of selected applications. CDNs may combine admission control and resource reservation, caching, traffic engineering, and rate control techniques to form a network architecture that supports their services. The goal is often to direct end users to the requested content in a manner that is responsive to network conditions, including congestion. Admission control and resource reservation may be used to provide a minimum throughput for selected content. Caching is used extensively in CDNs to direct users to a nearby server that has the requested content, to a relatively lightly loaded server, or to a server in an uncongested network. Traffic engineering can be implemented by CDNs using either routing (for facilities-based CDNs), geo-location methods that steer users to locally cached content based on the user's location, or a combination of the two. Rate control may be used to determine the speed at which specific content is served, to limit the speed at which content is served to an individual user, or to limit the speed at which content is served for a selected ASP.

6. Examples of Congestion Management Practices Based on User or Application

Section 5 discussed congestion management techniques. Section 6 illustrates congestion management practices that use these techniques.

As discussed in Section 4.1, this report uses the term congestion management practice to refer to the use by a party or organization:

- of a collection of traffic management techniques,
- targeted at particular users and/or applications,
- upon the trigger of some event.

As discussed above, network protocols that are application-agnostic may result in a QoE that is not acceptable to some applications. For example, packet routing is primarily configured to deliver traffic from source to destination, without regard to congestion. While delay-tolerant applications such as web browsing and email may tolerate changes in routes, delay-intolerant applications such as VoIP and gaming may not.

In addition, during emergencies, ISPs generally implement extreme congestion management practices that target specific user groups and applications. The goal of these practices is to give preference to first responder and government communications, voice calls (especially emergency calls) and low-bandwidth messages (e.g., SMS messages) originating inside the impacted area. Non-government calls and messages that originate outside the impacted area towards destinations inside the area are routinely restricted in order to ensure capacity is available for outbound calls. Traffic that cannot be readily recognized as voice or low-bandwidth messaging is also restricted.

The congestion management techniques discussed in Section 5 may be configured based on user or application criteria. Unfortunately, no single technique works in every situation. The practices discussed below illustrate how network operators may combine user- or application- based congestion management techniques.

Network operators face tradeoffs when designing congestion management practices. Application- or user- based congestion management practices may achieve better performance for selected applications. Such practices also may enable service providers to offer connectivity products that cater to particular customer’s tastes or needs. However, they add complexity, which may result in added costs that each network operator will evaluate. In some cases they may be harmful to applications.

The offerings of a service provider often include consumer broadband Internet access, enterprise broadband Internet access, and other IP-based services. Multiple services often utilize the same network links and routers, and the benefit of sharing links and routers among multiple services emanates from the fundamental design principle of the Internet. As discussed in Section 3.1, the Internet is designed so that multiple users may dynamically share capacity. The statistical multiplexing of traffic between different services results in a design that is more efficient than the offering of different services using separate links and routers.

However, sharing capacity between multiple services may be facilitated by the use of a number of congestion management techniques described above. For instance, the sharing of capacity between prioritized video, prioritized voice, and best-effort Internet traffic may use packet scheduling techniques (Section 5.7) that can prevent best-effort Internet traffic from degrading the quality of the prioritized video and voice traffic. In addition, network operators may also limit the amount of prioritized voice and video traffic using admission control techniques, as discussed in Section 5.2.

Network operator practices determine the capacity allocation to each service. In some cases, capacity may be permanently divided between multiple services. In other cases, capacity might be reserved for some services, but available to other services when not in use. For example, a network operator may reserve capacity for prioritized voice and video traffic. If a network operator uses admission control techniques on these services, then it may limit the impact of congestion on the QoE of these services while allowing degradation to the QoE of other services. However, when there is insufficient demand for the prioritized voice and video services, the reserved capacity may be available to the other services.

6.1. TCP Connection Termination Practices for Control of Peer-to-Peer Traffic

In the early-to-mid 2000s, some ISPs began to experience congestion as a result of the increased popularity of peer-to-peer (P2P) file-sharing. Specifically, one identified cause of congestion was users seeding P2P networks with files for others to download. In this situation, a user's P2P application would make heavy use of that user's or a neighborhood's upstream network link, often causing congestion, particularly on cable networks with upstream access links multiplexed among numerous users in a neighborhood [Martin and Westall]. This congestion may have had detrimental effects on the performance of delay-sensitive applications sharing the upstream connection (VoIP, for example).

One solution to this congestion management problem that gained traction among some ISPs involved deploying deep packet inspection (DPI) technology to limit the number of upstream TCP connections for peer-to-peer applications at particular points on the network. In a typical implementation, the equipment was configured to identify peer-to-peer application protocols associated with the heaviest traffic usage, such as BitTorrent [Comcast 2008]. Thus, this practice was application-based. When the number of upstream-only TCP connections associated with any of the P2P protocols reached a pre-defined threshold in a particular location on the network (indicating potential congestion in the upstream connection), the device would send a TCP "reset" packet to both sides of the TCP connection for the P2P file exchange, causing the TCP connection between the two peers to be closed and the exchange to cease.

The connection termination approach was based on the premise that most P2P clients are designed to automatically re-start a download after an interruption (assuming another peer elsewhere has made the same file or file portion available). Assuming that the downloading peer would re-start by connecting to a peer on a different ISP or in a less congested part of the ISP's network, the connection termination approach would have the effect of moving upstream P2P traffic out of the congested upstream link and reducing the overall traffic utilization on the upstream link while introducing only a small delay in the time it took for the downloader to finish the file transfer. Uploaders were assumed to be mostly indifferent as to whether their file uploads were interrupted or how quickly they finished.

However, in the case where a file is seeded by only a single peer, the connection termination approach could effectively prevent a downloader from obtaining the file. This scenario was verified and reported in the press in 2007 [Svensson], leading to a proceeding at the FCC [FCC 2008] and causing ISPs to shift away from using TCP connection termination.

Application-based connection termination can be problematic not only because it can prevent access to specific content in some cases, but also because the precise identification of traffic associated with particular applications can be difficult. As explained in BITAG's Port Blocking report, many applications can share the same port number, making it difficult to identify applications solely based on the ports they use [BITAG Port Blocking Report]. Even when DPI is employed, many applications have developed sophisticated designs to help their traffic blend in with other common traffic that is unlikely to be the target for termination, such as web traffic. Incorrectly identifying traffic causes connections to be unintentionally terminated.

Application-based connection termination also potentially creates fragile dependencies between application behavior and traffic management practices. The system described above relied on the fact that many P2P clients would automatically seek out a new peer from which to download if a connection to an existing peer were terminated. With the system in place, the designers of P2P applications would need to maintain that re-connection behavior in order for their applications to continue to work under the broadest set of circumstances, even if choosing a different design that did not involve automatic re-connection would be preferable for other reasons. From an application design perspective, it is suboptimal for the traffic management practices used on individual networks to dictate design choices for applications that are meant to be used on any network.

Alternative congestion management practices to cope with P2P traffic have since been proposed [ALTO] [DECADE] [HOMENET] [RFC 5594] [RFC6789] [RFC6817].

6.2. Traffic Shaping Practices

As discussed in Section 5.4, a network operator may use traffic shaping on the basis of user or application to control the rate of selected flows. The goal may be to limit the extent selected users contribute to congestion based on their behavior, to limit the extent selected applications contribute to congestion, or to allocate additional capacity for a limited time to selected flows. Traffic shaping on the basis of user or application is different than the reduction of capacity that users may experience at peak hours as a by-product of heavy load or congestion on the shared portions of the network. A network operator using traffic shaping determines which flows are subject to traffic shaping and the corresponding capacity limits as part of a traffic management practice. There are three essential elements to a practice: when the technique is applied, to which flows it is applied, and the capacity limits imposed.

If the goal is to limit the extent to which selected users contribute to congestion based on their behavior, a network operator may apply traffic shaping when it expects recurrent congestion or when it measures congestion. For example, the UK ISP Virgin Media implements traffic shaping during times of the day when recurrent congestion is known to occur [Virgin Guide to Traffic Management]. In contrast, a network operator may measure congestion in different parts of its network, and apply traffic shaping only when congestion rises above a threshold, and only in the part of the network experiencing congestion. When a network operator decides to use traffic shaping, it must select the flows to which it should be applied. For example, Virgin Media applies traffic shaping to users who have transferred an amount of data above a threshold during the peak period. Finally, the network operator must decide which capacity limit to impose upon those flows. For example, an ISP may temporarily reduce the maximum upstream or downstream transmission rate of a targeted user.

If the goal is to limit the extent to which selected applications contribute to congestion, a network operator may apply traffic shaping on the basis of application criteria. For example, in some of Virgin Media's broadband Internet packages, capacity limits are placed on P2P and newsgroup applications during peak periods.

If the goal is to allocate additional capacity for a limited time to selected flows, a network operator may apply traffic shaping on the basis of user. For example, PowerBoost™, formerly offered by several US ISPs, increased the capacity available to flows during transmission of the initial portion of the flow [Comcast Powerboost] [End-to-end Detection] [Measuring Broadband America 2012] [Powerboost]. This practice was applied without regard to whether congestion was occurring at that time in that portion of the network, and was targeted at all subscribers of certain broadband Internet services.

The impact of a traffic shaping practice upon users and applications depends on how narrowly tailored the practice is. Practices that are applied during time periods when recurrent congestion is known to occur will impact more users than practices that are applied only when congestion is detected. Practices that are applied to all users will impact more users than practices that are applied only in the part of the network experiencing congestion or only to heavy users.

6.3. Prioritization Practices to Handle Heavy Users

As discussed in Section 5.7, a network operator may use packet queuing and scheduling techniques to allocate what it sees as a fair share of resources between users. The goal may be to limit the extent to which selected users contribute to congestion based on their behavior. Packet scheduling may thus be used as an alternative to traffic shaping. As with traffic shaping practices, there are three essential elements to a prioritization practice: when the technique is applied, to which flows it is applied, and the scheduling actions taken.

In the Virgin Media practice that limits heavy users, described in Section 6.2, the congestion management practice is applied during peak usage periods to users who have transferred an amount of data above a certain threshold level during the peak period. An alternative method for managing the impact of heavy users makes use of packet scheduling techniques rather than traffic shaping techniques. An example of this approach is the FairShare congestion management practice developed by Comcast [RFC6057]. Rather than being applied during peak usage periods that last for hours, the FairShare practice is only applied when the network load rises above a threshold just below the traffic level at which congestion would occur. An alternative version, Sandvine Fairshare, is only applied when the delay in the access network rises above a threshold [Sandvine 2013]. Additionally, Comcast FairShare only targets those users who are in regions of the network that are congested and who have transmitted at an average rate above a certain threshold percentage of their service tier rate (e.g. 70%) during the most recent measurement interval (e.g. 15 minutes), rather than targeting all users who have transferred an amount of data above a threshold.

Traffic shaping practices intended to handle heavy users temporarily reduce the maximum upstream or downstream transmission rate of a targeted user. In contrast, prioritization practices use packet scheduling techniques to change the transmission ordering or priority of the packets. All traffic is initially labeled with the same priority classification. When the practice is applied, it temporarily lowers the priority of packets to or from users who are disproportionately contributing to the congestion. This type of congestion management practice is user-based because it targets heavy users, but not application-based because packets are classified without regard to the application that generated them. If congestion does not occur, this lower priority level may have no effect upon packet transmissions. When congestion does occur, packets with the lower priority level will experience higher delay and perhaps higher loss than packets with the higher priority level. For example, in the FairShare congestion management practice lower priority packets are transmitted only if residual capacity remains after high-priority packets are transmitted.

6.4. Transcoding Practices

As discussed in Section 5.4, one form of rate control by ISPs is transcoding, in which content compressed at the source is decoded and re-encoded by a network device into a flow at a lower rate. This technique may be applied by an ISP when a flow is progressing from a network with a higher capacity to a network with a lower capacity, which often occurs when traffic originates on a broadband wired network and terminates at a device on a cellular network since cellular network capacity is often lower than that of wired links. The goals of a transcoding practice are to avoid congestion, to increase performance through faster downloads, and to reduce consumer costs through lower data usage. As with other congestion management practices, there are three essential elements to a transcoding practice: when transcoding is applied, to which flows it is applied, and the type of re-encoding done.

The trigger for when transcoding is applied by an ISP can vary. If the goal is to decrease download times or to reduce data usage, then a network operator may use transcoding across its entire network at all times. If the goal is to avoid congestion, then a network operator may use transcoding only when and where congestion is present.

Transcoding may be applied to content that has been compressed at the source. Compression encodes information using fewer bits than the original representation. Lossless compression reduces bits by identifying and eliminating statistical redundancy, and thus results in no degradation in the quality of the information. Lossy compression reduces bits in a manner that results in minimal degradation of quality of the information. Compression is often applied at the source to audio and video. Some sources compress audio and video for transmission at a variety of rates, while other sources only compress at a single rate. Transcoding by an ISP may be applied to flows that an ISP detects to be audio or video, in a format that the ISP recognizes and has the ability to decode and re-encode. For example, Verizon Wireless may apply transcoding to all content files (primarily video) originating from Internet webservers using common compression formats [Verizon Optimization Deployment]. Other content, including audio and video flows that have been encrypted, cannot be easily transcoded.

When transcoding is applied, targeted flows are decoded and re-encoded using a compression algorithm that reduces the file size. The choice of the compression algorithm affects both the reduction in file size and the potential degradation in quality. Transcoding that uses heavy compression can lead to larger "savings" when it comes to bandwidth, but also raises the likelihood of significantly, and importantly noticeably, affecting the quality of the audio or video. The QoE of an audio or video stream depends on both the QoS characteristics of the stream and the capabilities of the device rendering the audio and video. In cellular networks, many devices have screens that are smaller and of lower resolution than a computer monitor. For such devices, the decrease in QoE resulting from transcoding using moderate compression may not be noticeable to the user. However, for devices such as tablets or laptops with high resolution screens, the decrease in QoE resulting from transcoding using moderate compression may be noticeable, for example resulting in reduced color accuracy and sharpness. An ISP that uses transcoding across a large portion of its network at all times may thus be more inclined to use only light compression in order to minimize the impact on all users. An ISP that uses transcoding only when and where congestion is present may adjust the compression level based upon current network conditions and use heavier compression when conditions warrant.

A transcoding practice is application-based because it is applied only to particular file formats identified by the network operator. It may also be a user-based congestion management practice if a network operator only transcodes files transmitted to or from specific users. It is also possible to apply transcoding based on the receiving device by selecting the transcoding algorithm on the basis of the device's capabilities. If applied in any of these manners, transcoding is not based on content itself nor on the content originator. For example, Verizon Wireless's transcoding practice depends on the compression format of the content file, but does not depend on the content of the file or the originating web site.

There are substantial tradeoffs in the use of transcoding practices. The QoE of a video depends both on the QoS characteristics of the flow and on the compression level. Poor QoS metrics may result in stuttering, pixilation, or freezing. Thus, the QoE of a transcoded video may be higher than that of a non-transcoded video if the benefit resulting from decreased congestion outweighs the noticeable degradation resulting from higher levels of compression. However, if the source compresses audio and video for transmission at a variety of rates, and if it is capable of dynamically selecting the compression rate in response to the user's device or network congestion, then it may be preferable for the source to dynamically select the encoding rather than for an ISP to transcode. The optimal solution may depend on whether the content provider or the ISP has better information about a device's capabilities, congestion, and user preferences.

6.5. Resource Reservation Practices to Improve the Performance of Applications Needing Minimum Bandwidth

A congestion management practice that can effectively improve the QoE of applications that have minimum bandwidth requirements is to reserve resources specifically for a set of traffic flows, as discussed in Section 5.2.

One way to reserve resources is to dedicate certain frequencies for specific traffic. For example, consumers using dual-band Wi-Fi routers may reserve one of the frequency bands for specific traffic. Cable providers also reserve resources for broadcast channels through dedicated frequencies.

Another way to accomplish a similar effect is to limit other traffic to a maximum capacity. For example, consumers may configure home routers to place a limit on the upstream capacity used by certain flows. Using this congestion management practice, only selected traffic may use the reserved capacity. As discussed in Section 5.1, classification of packets may be performed using attributes from multiple layers, including protocol codepoints, and source or destination addresses. For example, a video streaming application may apply a classification marking to its video packets to indicate a request for reserved capacity. Alternatively, the consumer may configure CPE to apply a classification marking to all packets from a chosen application to indicate a request for reserved capacity. An enterprise may classify all traffic from addresses of VoIP telephones to indicate a request for reserved capacity. An ISP may reserve capacity for all traffic between different locations of an enterprise customer.

This congestion management practice can be used to ensure that a large volume of traffic using the unreserved resources does not negatively impact the flows using the reserved resources. However, this practice restricts the capacity made available to traffic using unreserved resources, which may cause that traffic to more frequently experience congestion than if all capacity were allocated to unreserved use.

When a network operator offers resource reservation, it is important that it limit the percentage of capacity that can be reserved based on the desired QoE of applications using the reserved capacity. For example, Cisco recommends that no more than 33% of traffic be allowed into a real-time Expedited Forwarding queue [Cisco QoS Design Overview]. Because the capacity that can be reserved is fixed, it is important that traffic using the reserved capacity be limited to the reservations made, as discussed in Section 5.2. When a source transmits more traffic than the capacity it has reserved, congestion management practices will generally apply policing techniques to either drop the packets that exceed the reserved capacity or reclassify them to use the unreserved resources. Since this will generally have a negative impact on the QoE of the applications expecting to use reserved capacity, and since the reserved capacity is often considerably smaller than the unreserved capacity, it is frequently preferable for all of an application's traffic to use the unreserved capacity than to exceed reserved capacity.

Techniques that allow per-flow reservation (including IntServ described in Section 5.8.1, 3GPP's Policy and Charging Control described in Section 5.8.5, and PacketCable DOCSIS QoS described in Section 5.8.3) are less susceptible to the dangers of traffic exceeding reserved capacity than are techniques that rely solely on static packet classification. Network operators that use static packet classification to admit traffic into reserved resources tend to tightly control the maximum traffic that can be classified in this manner. IPTV providers do this by restricting the maximum number of HD and SD audio and video flows that can access reserved capacity in contracts with their customers. VoIP providers often do this by restricting the maximum number of simultaneous calls that can access reserved capacity.

As discussed in Section 5.2, the offering and granting of resource reservations are decisions made by each network operator. Reservations within an ISP's network thus typically require a business arrangement between the ISP and a consumer, enterprise, ASP, or another ISP, e.g. as part of an agreement to create an intranet or virtual private network (VPN) between different locations of an enterprise customer.

6.6. Preferential Treatment Practices to Improve the Performance of Delay-and Loss-Intolerant Applications

As discussed in Section 3.5, delay-intolerant applications, such as VoIP, are likely to have a QoE that is very sensitive to end-to-end delay. Such applications will thus usually not request that dropped packets be retransmitted from the sender and will usually discard packets that do not arrive within a certain time interval. Even brief occurrences of congestion, on the order of a few tenths of a second, can cause noticeable degradation in the QoE of these applications. As discussed in Sections 3.1-3.2, the Internet is not designed to avoid all such brief occurrences of congestion. Loss-intolerant applications (e.g., streaming video) are likely to have a QoE that is sensitive to lost packets. Requesting retransmission of these packets increases jitter, which these applications also do not tolerate well.

An effective congestion management practice is to implement packet scheduling and packet dropping techniques in a manner that prioritizes delay-intolerant or loss-intolerant applications. One way to prioritize selected traffic is to place it into a separate queue that is served at a preferential rate compared to other traffic, or that has a "do not drop" policy, as discussed in Section 5.7. For example, a consumer may configure CPE so that all packets from VoIP applications are classified as high priority for low delay. An enterprise may classify all traffic from VoIP telephones as high priority for low delay. An ISP may classify VoIP packets from its own telephone customers as high priority for low delay.

Another way to prioritize traffic is to combine packet scheduling, packet dropping, and admission control techniques to provide different classes of service, as discussed in Sections 5.8.1-5.8.2. For example, an ASP or enterprise may classify VoIP packets with a DiffServ codepoint intended for VoIP.

Using this congestion management practice, congestion management techniques such as packet scheduling are based in part on the classification of a packet. As discussed in Section 5.1, classification of packets may be performed using attributes from multiple layers, including protocol codepoints, and source or destination addresses. For example, a VoIP application may apply a classification marking to its packets to indicate a request for high priority classification. As discussed in Section 6.5, it is very important to ensure that packets that can receive such classification are limited (through contracts, agreements, policing techniques, or by the devices initiating the traffic) to maximum rates of traffic. If it is not possible to ensure such limits, it is generally preferable to apply no preferential treatment.

This congestion management practice can be used to ensure that congestion has only a limited negative impact upon selected flows, providing that high-priority traffic is strictly limited. However, as previously mentioned, this practice restricts the capacity made available to lower priority traffic, which may cause that traffic to more frequently experience congestion than if no prioritizations were used.

As discussed in Sections 5.8.1-5.8.2, whether to use prioritization and preferential treatment of packets are decisions made by each network operator. The honoring of prioritizations requested by other parties thus typically require a business arrangement between the ISP and a consumer, enterprise, ASP, or another ISP, e.g., as part of an agreement to create an intranet between different locations of an enterprise customer.

7. Technical Working Group (TWG) Recommendations

This section of the report presents recommendations of the BITAG Technical Working Group (TWG). These recommendations include transparency concerning congestion management practices, using standardized or industry-reviewed congestion management practices, minimizing impact to users or applications, encouraging congestion management based on users' expressed preferences, testing of application-based criteria, and using efficient and adaptive network resource management.

The BITAG recognizes that there may be times when immediate and acute security issues or operational issues may preclude the use of suggested practices for some period of time.

7.1. Transparency

The BITAG suggests that ISPs and ASPs should disclose information about their user- or application- based network management and congestion management practices for Internet services such that the information is readily accessible to the general public. This information should be made available on network operators' public web sites and through other typically used communications channels, including mobile apps, contract language, or email. ISPs and ASPs may choose to use a layered notice approach, using a simple, concise disclosure that includes key details of interest to consumers complemented by a more thorough and detailed disclosure for use by more sophisticated users, application developers, and other interested parties. The detailed disclosure should include:

- descriptions of the practices;
- the purposes served by the practices;
- what types of traffic are subject to the practices, if not all traffic, and, if appropriate, a general explanation of how traffic types are identified on the network;
- the practices' likely effects on end users' experiences;
- the triggers that activate the use of the practices and whether those triggers are user- or application- based;
- the approximate times at which the practices are used, if they are limited to particular times of day; and
- which subset of users may be affected, if not all, e.g. the geographic boundaries in which the practices are used.

The disclosure should also include the predictable impact, if any, of a user's other subscribed network services on the performance and capacity of that user's broadband Internet access services during times of congestion, where applicable. These disclosures should complement ISPs' other disclosures that describe their Internet service offerings, expected and actual access speeds, performance characteristics, and suitability for supporting particular kinds of applications.

All of these disclosures may be made without divulging competitively sensitive information and should be consistent with the requirements described in the FCC's *Open Internet Order* [Open Internet Order].

7.2. Network Operators should use accepted industry "Best Practices," standardized practices, or seek industry review of practices.

Organizations like NANOG and the IETF produce considered recommendations of Best Practices and standard practices for a variety of operational issues, including congestion and congestion management. One example is RFC 2309, which recommends that network queues be managed using one of a class of "Active Queue Management" algorithms in order to maximize average throughput while minimizing average latency during transient congestion events without unnecessarily targeting individual data flows or applications. Where network operators see the need for an innovative solution that has not been standardized or documented as a Best Practice, they should consider bringing their unique network or congestion management practices to groups such as the IETF, NANOG, or other technical industry bodies for discussion and possible documentation. An example of this is RFC 6057, an informational RFC entitled "Comcast's Protocol-Agnostic Congestion Management System". This recommendation for review is not intended to imply that the network operator wait until completion of the review before implementing the practice.

7.3. When engaging in a congestion management practice that could have a detrimental impact on the traffic of certain users or certain applications, the practice should be designed to minimize that impact.

Some congestion management practices may cause certain users or certain applications to experience performance degradation (see the practices in Sections 6.2 and 6.3, for example). ISPs and ASPs should seek to minimize the degradation of certain users or applications, to the extent possible, while still managing the effects of the congestion that triggered the use of the practice. For example, if a particular congestion issue occurs only at certain locations within the network or during certain times of day, practices that degrade the experience for some users or applications should focus on those locations and times, rather than being applied at all locations and times.

The practice described in RFC 6057 provides an example of this minimization. This practice has the potential to deprioritize some users' traffic, but its application is limited to a minimal subset of heavy users. The practice also only targets those users who are in regions of the network that are congested and only lasts until utilization returns to a normal level. Thus, although some heavy users' performance may be degraded, that degradation is limited in terms of the number of users affected, the locations in the network where users can be affected, and the amount of time during which they are affected.

7.4. If application-based congestion management practices are used, those based on a user's expressed preferences are preferred over those that are not.

User- and application- agnostic congestion management practices are useful in a wide variety of situations, and may be sufficient to accommodate the congestion management needs of network operators in the majority of situations. However, at times network operators may choose to use application-based congestion management practices. For example, congestion management practices based on application criteria may allow an ISP to better honor customer preferences, such as a preference for prioritized VoIP.

The BITAG recommends that if application-based congestion management practices are used, those that prioritize application traffic according to a user's expressed preferences be preferred over those that do not.

7.5. If application–based criteria are used by a network operator, they should be tested prior to deployment and on an ongoing basis.

Application-based classification by network operators (e.g., using DPI) can sometimes be erroneous. If network operators choose to use application-based criteria for congestion management, the accuracy of the application traffic classifier should be tested before deployment and on an ongoing basis thereafter to ensure that applications are not misclassified and thus mistakenly affected by congestion management.

7.6. ASPs and CDNs should implement efficient and adaptive network resource management practices.

The BITAG recommends that ASPs and CDNs match use of network resources to the performance requirements of the application. Inefficient or non-adaptive application design may lead to traffic that is more susceptible to degradation by congestion management techniques, to the detriment of the application and its users.

First, applications should be designed to efficiently use network resources, e.g. by using efficient compression or by basing their transmission rate on the capabilities of the receiving device. ASPs and CDNs should recognize that capacity is shared and limited, and that use of network resources by one application reduces the capacity available to other applications and other users.

Second, applications should be designed to adaptively use network resources to the extent feasible given the application's requirements. As discussed in Section 3.5, the characteristics of an application determine when and how QoE is degraded by congestion. Applications should be designed to adaptively adjust their transmission rates based on any congestion in the network. The period over which they adapt should be based on the application's requirements. Jitter-intolerant applications should at a minimum adapt within a period of time equal to their buffering capability. Applications tolerant of variations in throughput over time periods of seconds or greater should adapt within a period of seconds or greater, e.g. by using the TCP protocol.

8. References

[802.1p] The Institute of Electrical and Electronics Engineers, Inc., “IEEE Standard for Local and Metropolitan Area Networks—Media access control (MAC) Bridges”, 2004, <http://standards.ieee.org/getieee802/download/802.1D-2004.pdf >.

[802.1Q] The Institute of Electrical and Electronics Engineers, Inc., “IEEE Standard for Local and metropolitan area networks--Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks”, 2011, <http://standards.ieee.org/getieee802/download/802.1Q-2011.pdf>.

[ALTO] Alimi, R., R. Penno, and Y. Yang, “ALTO Protocol”, Internet Draft, July 14, 2013, <http://datatracker.ietf.org/doc/draft-ietf-alto-protocol/>.

[3GPP PCC specification] 3GPP, “Policy and Charging control Architecture”, <http://www.3gpp.org/ftp/Specs/html-info/23203.htm>.

[Atias] Atias, Ronen, “Under the Hood of the DDoS Attack on U.S. Banks”, Incapsula, Blog Post, January 8, 2013, <http://www.incapsula.com/the-incapsula-blog/item/603-cyber-attack-us-banks>.

[BGPMON] BGPMON, “Looking at the spamhaus DDOS from a BGP perspective”, March 30, 2013, <http://www.bgpmon.net/looking-at-the-spamhouse-ddos-from-a-bgp-perspective/>.

[BITAG Port Blocking Report] BITAG Technical Working Group, “Port Blocking”, August 2013, <http://www.bitag.org/documents/Port-Blocking.pdf>.

[Cisco QoS Design Overview] Cisco, “Quality of Service Design Overview”, <http://www.cisco.com/en/US/docs/solutions/Enterprise/WAN_and_MAN/QoS_SRND/QoSIntro.html#wp46447>.

[CiscoVNI] Cisco, “VNI Forecast Highlights”, 2012, <http://www.cisco.com/web/solutions/sp/vni/vni_forecast_highlights/index.html>.

[Comcast2008] Comcast Corporation, Comcast 2008 Network Management Practices,
<http://downloads.comcast.net/docs/Cover_Letter.pdf>,
<http://downloads.comcast.net/docs/Attachment_A_Current_Practices.pdf>,
<http://downloads.comcast.net/docs/Attachment_B_Future_Practices.pdf>,
<http://downloads.comcast.net/docs/Attachment_C_Compliance_Plan.pdf>,

<http://downloads.comcast.net/docs/comcast-nm-transition-notification.pdf>, <http://downloads.comcast.net/docs/january-30-2009-comcast-fcc-response.pdf>.

[Comcast Powerboost] Woundy, R., "Powerboost: A Comcast Innovation for High-Speed Internet Users", Aug. 24, 2011, <http://corporate.comcast.com/comcast-voices/powerboost-a-comcast-innovation-for-high-speed_internet_users>.

[DECADE] Alimi, R., A Rahman, D. Kutscher, Y. Yang, H. Song, and K. Pentikousis, "DECADE: DECoupled Application Data Enroute", Internet Draft, June 10, 2013, <http://tools.ietf.org/html/draft-alimi-protocol-01>.

[End-to-end Detection] Kanuparthy, P., and C. Dovrolis, "End-to-end Detection of ISP Traffic Shaping Using Active and Passive Methods", Georgia Institute of Technology, <http://www.cc.gatech.edu/~partha/shaperprobe-TR.pdf>.

[Estonia] Richards, J., "Denial-of-Service: The Estonian Cyberwar and Its Implications for U.S. National Security", International Affairs Review, <http://www.iar-gwu.org/node/65>.

[Evolution of Internet Congestion] Bauer, S., D. Clark, and W. Lehr, "The Evolution of Internet Congestion", 2009, <http://people.csail.mit.edu/wlehr/Lehr-Papers_files/Bauer_Clark_Lehr_2009.pdf>.

[FCC 07-31] Federal Communications Commission, Broadband Market Practices Notice of Inquiry, FCC 07-3, 2007.

[FCC 2008] Federal Communications Commission, Memorandum Opinion and Order In the Matters of Free Press and Public Knowledge Against Comcast Corporation for Secretly Degrading Peer-to-Peer Applications; Broadband Industry Practices; Petition of Free Press Et Al. for Declaratory Ruling That Degrading an Internet Application Violates the FCC's Internet Policy Statement and Does Not Meet an Exception for "Reasonable Network Management.", 2008.

[GSMA] GSM Association, "Advancing 3GPP Networks: Optimisation and Overload Management Techniques to Support Smart Phones, Version 1.0", White Paper, June 2012, <http://www.gsma.com/newsroom/wp-content/uploads/2013/02/Opt-Overload-Man-Tech-Support-Smart-Phones.pdf>.

[Guardian] Shoffield, J., "Slammer Worm Broke Net Speed Records", The Guardian, Technology Blog, Feb. 6, 2003, <http://www.guardian.co.uk/technology/blog/2003/feb/06/slammerwormbr>.

[HOMENET] Chown, T., J. Arkko, A. Brandt, O. Troan, and J. Well, "Home Networking Architecture for IPv6", Internet Draft, Aug. 1, 2013, <http://datatracker.ietf.org/doc/draft-ietf-homenet-arch/>.

[Kamphuis] Kar, I., "Meet the Man Behind the Biggest Cyberattack in History", Heavy.com, Tech Interview, March 29, 2013, <http://www.heavy.com/tech/2013/03/sven-olaf-kamphuis-cyberbunker-stophaus-spamhaus-cyberattack/>.

[Kurose and Ross] Kurose, James F., and Ross, Keith W., "Computer Networking: A Top-Down Approach", Addison-Wesley, 6th edition, 2013.

[Martin and Westall] Martin, J., and J, Westall, "Assessing the Impact of BitTorrent on DOCSIS Networks", Proceedings of the 2007 IEEE Broadnets, 2007.

[Measuring Broadband America 2012] Federal Communications Commission, Office of Engineering and Technology and Consumer and Governmental Affairs Bureau, "2012 Measuring Broadband America, July Report", July 2012, <http://transition.fcc.gov/cgb/measuringbroadbandreport/2012/Measuring-Broadband-America.pdf>.

[Open Internet Order] Federal Communications Commission, In the Matter of Preserving the Open Internet Broadband Industry Practices, FCC 10-201, Dec. 21, 20120, <http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-10-201A1.pdf>.

[PacketCable] Cablelabs, "PacketCable", <http://www.cablelabs.com/packetcable/>.

[Powerboost] Bauer, S., D. Clark, W. Lehr, "Powerboost", <http://groups.csail.mit.edu/ana/Publications/p7-bauer.pdf>.

[RFC1122] R. Braden, "Requirements for Internet Hosts – Communications Layers", October 1989, RFC 1122, <http://tools.ietf.org/html/rfc1122>.

[RFC 1633] Braden, R., D. Clark, and S. Shenker, "Integrated Services in Internet Architecture: An Overview", RFC 1633, June 1994, <http://tools.ietf.org/html/rfc1633>.

[RFC2205] Braden, R., L. Shang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP)", RFC 2205, Sept. 1997, <http://tools.ietf.org/html/rfc1633>.

[RFC2309] Braden, B., D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclawski, and L. Zhang. "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC, 2309, April 1998, <http://tools.ietf.org/html/rfc2309>.

[RFC2430] Li, T., and Y. Rekhter, "A Provider Architecture for Differentiated Services and Traffic Engineering (PASTE)", RFC 2430, Oct. 1998, <http://tools.ietf.org/html/rfc2430>.

[RFC2474] Nichols, K., S. Blake, F. Baker, and D. Black, "Definition of the Differentiated

Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, Dec. 1998, <http://tools.ietf.org/html/rfc2474>.

[RFC2475] Blake, S., D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, Dec. 1998, <http://tools.ietf.org/html/rfc2475>.

[RFC2597] Heinanen, J., T. Finland, F. Baker, W. Weiss, J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999, <http://tools.ietf.org/html/rfc2597>.

[RFC3168] Ramakrishnan, K., S. Floyd, and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001, <http://tools.ietf.org/html/rfc3168>.

[RFC3246] Davie, B., A. Charny, J.C.R. Bennett, K. Benson, J.Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu, D. Stiliadis, "An Expedited Forwarding PHB (Per Hop Behavior)", RFC 3246, March 2002, <http://tools.ietf.org/html/rfc3246>.

[RFC3272] Awduche, D., A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "Overview and Principles of Internet Traffic Engineering", RFC 3272, May 2002, <http://tools.ietf.org/html/rfc3272>.

[RFC3550] Schulzrinne, H., S. Casner, R. Frederick, and V. Jacobsen, "RTP: A Transport Protocol for Real-Time Applications", RFC 3550, July 2003, <http://tools.ietf.org/html/rfc3550>.

[RFC4594] Babiarz, J., K. Chan, and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, Aug. 2006, <http://tools.ietf.org/html/rfc4594>.

[RFC4915] Psenak, P., S. Mirtorabi, A. Roy, L. Nguyen, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007, <http://tools.ietf.org/html/rfc4915>.

[RFC5472] Zseby, T., E. Boschi, N. Brownlee, B. Claise, "IP Flow Information Export (IPFIX) Applicability", RFC 5472, March 2009, < http://tools.ietf.org/html/rfc5472>.

[RFC5594] Peterson, J., and A. Cooper, "Report from the IETF Workshop on Peer-to-Peer (P2P) Infrastructure, May 28, 2008", RFC 5594, July 2009, <http://tools.ietf.org/html/rfc5594>.

[RFC6057] C. Bastian, T. Klieber, J. Livingood, J. Mills, and R. Woundy, "Comcast's Protocol-Agnostic Congestion Management System", Dec. 2010, RFC 6057, Informational, <http://tools.ietf.org/html/rfc6057>.

[RFC6437] Amante, S., B. Carpenter, S. Jiang, and J. Rajahalme, "IPv6 Flow Label Specification", RFC 6437, Nov. 2011, <http://tools.ietf.org/html/rfc6437>.

[RFC6679] Westerlund, M., I. Johansson, C. Perkins, P. O'Hanlon, and K. Carlberg, "Explicit Congestion Notification (ECN) for RTP over UDP", RFC 6679, August 2012, <http://tools.ietf.org/html/rfc6679>.

[RFC6817] Shalunov, S., G. Hazel, J. Iyengar, and M. Kuehlewind, "Low Extra Delay Background Transport (LEDBAT)", RFC 6817, Dec. 2012, <http://tools.ietf.org/html/rfc6817/>.

[RFC6789] Briscoe, B, R. Woundy, and A Cooper, "Congestion Exposure (ConEx) Concepts and Use Cases", RFC 6789, December 2012, <http://tools.ietf.org/html/rfc6789>.

[Sandvine2013] Sandvine Incorporated, LLC. . "Global Internet Phenomena Report: 1H 2013", July 16, 2013, <http://www.sandvine.com/downloads/documents/Phenomena_1H_2013/Sandvine_Global_Internet_Phenomena_Report_1H_2013.pdf>.

[Svensson] Svensson, P, "Comcast Actively Hinders Subscribers' File-sharing Traffic, AP Testing Shows." Associated Press, October 19, 2007.

[Verizon Optimization Deployment] Verizon Wireless, "Optimization Deployment – Terms and Conditions", Aug. 17, 2013, <http://support.verizonwireless.com/support/information/network_optimization.html>.

[Virgin Guide to Traffic Management] Virgin Media, "Your Guide To Traffic Management", August 17, 2013, <http://my.virginmedia.com/traffic-management/traffic-management-policy-thresholds.html>.

9. Glossary of terms

All definitions of terms are solely for the purposes of this report. Readers should be aware that many terms have alternate definitions, particularly when used in different or non-networking contexts.

admission control	Decisions that determine which flows are granted reservations.
application layer	The layer at which network applications and application-specific techniques reside, e.g. HTTP.
Application Service Provider (ASP)	A provider of applications used on broadband Internet access services.

best-effort service	A network service in which routers transmit packets in the order in which they are received, and without regard to the source, the application that generated them, or the resulting QoE.
bottleneck	The link or router in a network path where demand is highest relative to capacity.
bursty	A traffic flow is bursty if its volume changes rapidly in time.
cache	To temporarily store popular content in multiple locations.
capacity	The capacity of a link is the number of bits per second that it can transmit. The capacity of a router is the number of packets or bytes per second that it can transmit.
classifier	An entity that selects packets based on the content of packet headers or other attributes according to defined rules.
codepoint	Specific bits in the IP packet header that can be used to classify packets of differentiated services in order to provide desired routing, scheduling, and dropping treatment.
congestion	The effect upon network performance during time periods in which instantaneous demand exceeds capacity.
congestion management practice	The use by a party or organization of a collection of traffic management techniques, targeted at particular users or applications, upon the trigger of some event.
congestion management technique	A specific congestion management function that determines whether Internet traffic is transmitted or the rate at which traffic is transmitted, or that enables such functions in other techniques.
data link layer	The layer at which packet forwarding and multiple access techniques reside, e.g. Ethernet.
delay jitter	The variation in end-to-end delay between packets.
demand	The volume of traffic that is presented to a link at a given point in time, typically measured in bits per second.
differentiated service	A description of the overall treatment of a subset of a customer's traffic across either an operator's network, a combination of operators' networks, or end-to-end.
end-to-end packet delay	The time from the transmission of a packet at the source to its reception at the destination.

end-to-end packet loss	The proportion of packets that are transmitted by a source that do not arrive at the destination.
end-to-end throughput	The average number of bits per second that are received at the destination.
flow	A group of packets that share a common set of properties.
Internet Service Provider (ISP)	A provider of broadband Internet access service.
layer	An abstraction that hides the implementation details of a particular set of functionality.
load	The ratio of demand (averaged over a chosen measurement interval) to the capacity of a specific link or router.
mark	Specific bits of a packet header that indicate the classification of a packet or that indicate congestion.
measurement interval	The time period over which measurements are aggregated.
network layer	The layer at which packet routing techniques resides, e.g. IP.
network operator	An ISP, or an ASP that operates a network.
physical layer	The layer at which digital-to-analog and analog-to-digital conversion techniques reside.
policing	The dropping or reclassification of packets that exceed the maximum capacity allocated to a flow.
Quality of Experience (QoE)	The satisfaction of a user with the performance of an application.
Quality of Service (QoS)	The performance desired by or required by an application, commonly expressed in terms of end-to-end packet delay, end-to-end packet loss, delay jitter, and/or throughput.
rate control	Adaptive modification of the rate of a traffic flow based on congestion.
Service Level Agreement (SLA)	Economic and legal agreements between network operators or between users and network operators that delineate contractual aspects of the service, often including the upstream and downstream bit rates at the boundary between the operators' networks, the maximum delay across an operator's network, sometimes the maximum proportion of packets to be dropped or other QoS characteristics, and sometimes specifications of payments.

statistical multiplexing	Multiplexing is the process whereby a router merges multiple traffic flows onto a single link. Statistical multiplexing, used in the Internet, uses packet scheduling to adapt the capacity used by each traffic flow to the flow's volume, resulting in a non-deterministic allocation.
traffic shaping	Rate limiting of flows in a network.
transcoding	Decoding and re-encoding of content at a router into a flow with a lower rate.
transport layer	The layer at which end-to-end flow or rate control techniques resides, e.g. TCP or UDP.
usage caps	Limits on the maximum number of bytes transmitted per month.
utilization	See "load".

10. Document Contributors and Reviewers

- Wendy Aylsworth, *Warner Bros.*
- Fred Baker, *Cisco*
- William Check, *National Cable and Telecommunications Association*
- Alissa Cooper, *Center for Democracy & Technology*
- Andrew Dugan, *Level 3*
- Amie Elcan, *CenturyLink*
- Michael Fargano, *CenturyLink*
- Jeff Finkelstein, *Cox Communications*
- David Fullagar, *Netflix*
- Greg Gewickey, *Warner Bros.*
- Bill Goodman, *Verizon*
- Dale Hatfield
- Trace Hollifield, *Bright House Networks*
- Scott Jordan
- Kevin Kahn, *Intel*
- Joe Lawrence, *Level 3*
- Jason Livingood, *Comcast*
- Chris Morrow, *Google*
- Donald Smith, *CenturyLink*
- Barbara Stark, *AT&T*
- Jeff Swinton, *Verizon*
- Matthew Tooley, *National Cable and Telecommunications Association*
- Jason Weil, *Time Warner Cable*