Internet Engineering Task Force (IETF)
Request for Comments: 6057
Category: Informational
ISSN: 2070-1721

C. Bastian
T. Klieber
J. Livingood
J. Mills
R. Woundy
Comcast

December 2010

Comcast's Protocol-Agnostic Congestion Management System

Abstract

This document describes the congestion management system of Comcast Cable, a large cable broadband Internet Service Provider (ISP) in the U.S. Comcast completed deployment of this congestion management system on December 31, 2008.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6057.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction 2
2. Applicability to Other Types of Networks 3
3. Key Terminology 3
4. Historical Overview 7
5. Summary 8
6. Relationship between Managing Congestion and Adding Capacity 9
7. Implementation and Configuration 10
7.1. Thresholds for Determining When a CMTS Port Is in a Near Congestion State 14
7.2. Thresholds for Determining When a User Is in an Extended High Consumption State and for Release from That Classification 15
7.3. Effect of BE Quality of Service on Users' Broadband Experience 19
7.4. Equipment/Software Used and Location 21
8. Conclusion 23
9. Exceptional Network Utilization Considerations 23
10. Limitations of This Congestion Management System 24
11. Low Extra Delay Background Transport and Other Possibilities ..24
12. Security Considerations 24
13. Acknowledgements 25
14. Informative References 26

1. Introduction

Comcast Cable is a large broadband Internet Service Provider (ISP), based in the U.S., serving the majority of its customers via cable modem technology. During the late part of 2008, and completing on December 31, 2008, Comcast deployed a new congestion management system across its entire network. This new system was developed in response to dissatisfaction in the Internet community as well as complaints to the U.S. Federal Communications Commission (FCC) regarding Comcast's old system, which targeted specific peer-to-peer (P2P) applications. This new congestion management system is protocol-agnostic, meaning that it does not examine or impact specific user applications or network protocols, which is perceived as a more fair system for managing network resources at limited times when congestion may occur.

It is important for readers to note that congestion can occur in any IP network, and, when it does, packets can be delayed or dropped. As Bob Briscoe has pointed out on an IETF mailing list, some amount of packet loss can be normal and/or tolerable, noting "But a single TCP flow with a round trip time (RTT) of 80 ms can attain 50 Mbps with a loss fraction of 0.0013% (1 in ~74,000 packets) so there's no need to try to achieve loss figures much lower than this. And indeed, if

flows aren't bottlenecked elsewhere, TCP will drive the system until it gets such loss levels. If, instead, a customer is downloading five separate 10 Mbps TCP flows still with an 80-ms RTT, TCP will drive losses up to 1 in ~3,000, or 0.03%, and any lower loss rates won't be able to improve performance". As a result, applications and protocols have been designed to deal with the reality that congestion can occur in any IP network, the mechanics of which we explain in detail later in this document.

The purpose of this document is to describe how this example of a large-scale congestion management system functions. This is partially in response to questions from other ISPs as well as solution developers, who are interested in learning from and/or deploying similar systems in other networks. In addition, it is hoped that such a document may help inform new work in the IETF, in the hope that better systems and protocols may be possible in the future. Lastly, the authors wish to transparently and openly document this system, so that there could be no doubt about how the system functioned.

2. Applicability to Other Types of Networks

Several document reviewers and other IETF participants have pointed out that, though we refer to functional elements that are specific to a Data Over Cable Service Interface Specification (DOCSIS)-based network implementation, this type of congestion management system could be generally applied to nearly any type of network. Thus, it is important for readers to take note of this and take into consideration that this sort of protocol-agnostic congestion management system could certainly fit in a wide variety of network types and implementations.

3. Key Terminology

This section defines the key terms used in this document. Some terms below refer to elements of the Comcast network. As a result, it may be helpful to refer to Figure 1 (see Section 7) when reviewing some of these terms.

3.1. Cable Modem

A device located at the customer premise used to access the Comcast High Speed Internet (HSI) network. In some cases, the cable modem is owned by the customer, and in other cases it is owned by the cable operator. This device has an interface (i.e., someplace to plug in a cable) for connecting the coaxial cable provided by the cable company to the modem, as well as one or more interfaces for connecting the modem to a customer's PC or home gateway device (e.g., home gateway,

router, firewall, access point, etc.). In some cases, the cable modem function, i.e., the ability to access the Internet, is integrated into a home gateway device or Embedded Multimedia Terminal Adapter (eMTA). Once connected, the cable modem links the customer to the HSI network and ultimately the broader Internet.

3.2. Cable Modem Termination System (CMTS)

A piece of hardware located in a cable operator's local network (generally in a "headend", Section 3.10) that acts as the gateway to the Internet for cable modems in a particular geographic area. A simple way to think of the CMTS is as a router with interfaces on one side leading to the Internet and interfaces on the other connecting to Optical Nodes and then customers, in a so-called "last mile" network.

3.3. Cable Modem Termination System (CMTS) Port

Also referred to simply as a "port". A port is a physical interface on a device used to connect cables in order to connect with other devices for transferring information/data. An example of a physical port is a CMTS port. A CMTS has both upstream and downstream network interfaces to serve the local access network, which are referred to as upstream or downstream ports. A port generally serves a neighborhood of hundreds of homes. Over time, CMTS ports tend to serve fewer and fewer homes, as the network is segmented for capacity growth purposes. Prior to DOCSIS version 3, a single CMTS physical port was used for either transmitting or receiving data downstream or upstream to a given neighborhood. With DOCSIS version 3, and the channel bonding feature, multiple CMTS physical ports can be combined to create a virtual port. A CMTS is also briefly defined in Section 2.6 of [RFC3083].

3.4. Channel Bonding

A technique for combining multiple downstream and/or upstream channels to increase customers' download and/or upload speeds, respectively. Multiple channels from the Hybrid Fiber Coax (HFC) network (Section 3.11) can be bonded into a single virtual port (called a bonded group), which acts as a large single channel or port to provide increased speeds for customers. Channel bonding is a feature of Data Over Cable Service Interface Specification (DOCSIS) version 3, as described in [DOCSIS_MULPI].

3.5. Coaxial Cable (Coax)

A type of cable used by a cable operator to connect customer premise equipment (CPE) -- such as TVs, cable modems (including eMTAs), and Set Top Boxes -- to the HFC network. This cable may be used within the home as well as in segments of the "last mile" network running to a home or customer premise location. There are many grades of coaxial cable that are used for different purposes. Different types of coaxial cable are used for different purposes on the network.

3.6. Comcast High Speed Internet (HSI)

A service/product offered by Comcast for delivering Internet service over a broadband connection.

3.7. Customer Premise Equipment (CPE)

Any device that resides at the customer’s residence, connected to the Comcast network, whether controlled by Comcast or not.

3.8. Data Over Cable Service Interface Specification (DOCSIS)

A reference standard developed by CableLabs that specifies how components on cable networks need to be built to enable HSI service over an HFC network, as noted in [DOCSIS_CM2CPE], [DOCSIS_PHY], [DOCSIS_MULPI], [DOCSIS_SEC], and [DOCSIS_OSSI]. These standards define the specifications for the cable modem and the CMTS such that any DOCSIS-certified cable modem will work on any DOCSIS-certified CMTS, independent of the selected vendor. The interoperability of cable modems and CMTSs allows customers to purchase a DOCSIS-certified modem from a retail outlet and use it on their cable-networked home. All DOCSIS-related standards are available to the public at the CableLabs website, at http://www.cablelabs.com.

3.9. Downstream

Description of the direction in which a signal travels, in this case from the network to a user. Downstream traffic occurs when users are downloading something from the Internet, such as watching a web-based video, reading web pages, or downloading software updates.

3.10. Headend

A cable facility responsible for receiving TV signals for distribution over the HFC network to the end customers. This facility typically also houses one or more CMTSs. This is sometimes also called a "hub".

3.11. Hybrid Fiber Coax (HFC)

A network architecture used primarily by cable companies, comprised of fiber-optic and coaxial cables that currently deliver Voice, Video, and Internet services to customers, as defined in Section 1.2 of [DOCSIS_MULPI].

3.12. Internet Protocol Detail Record (IPDR)

Standardized technology for monitoring and/or recording subscribers' upstream and downstream Internet usage data based on their cable modem. The data is collected from the CMTS and sent to a server for further processing. Additional information is available at http://www.ipdr.org, as well as [IPDR_Standard] and [DOCSIS_IPDR].

3.13. Optical Node

A component of the HFC network generally located in customers' local neighborhoods that is used to convert the optical signals sent over fiber-optic cables to electrical signals that can be sent over coaxial cable to customers' cable modems, or vice versa. A fiber-optic cable connects the Optical Node, through distribution hubs, to the CMTS, and coaxial cable connects the Optical Node to customers' cable modems.

3.14. Provisioned Bandwidth

The peak speed associated with a tier of service purchased by a customer. For example, a customer with a 105 Mbps downstream and 10 Mbps upstream speed tier would be said to be provisioned with 105 Mbps of downstream bandwidth and 10 Mbps of upstream bandwidth. This is often referred to as 105/10 service in industry parlance.

The Provisioned Bandwidth is the speed that a customer's modem is configured (and the network is engineered) to deliver on a regular basis (which is not the same as a "Committed Information Rate" or a guaranteed rate). Internet speeds are generally a best effort service that are dependent on a number of variables, many of which are outside the control of an Internet Service Provider (ISP). In general, speeds do not typically exceed a customer's provisioned speed. Comcast, however, invented a technology called "PowerBoost" [PowerBoost_Specification] that, for example, enables users to experience brief boosts above their provisioned speeds while they transfer large files over the Internet, by utilizing excess capacity that may be available in the network at that time.

3.15. Quality of Service (QoS)

A set of techniques to manage network resources to ensure a level of performance to specific data flows, as described in [RFC1633] and [RFC2475]. One method for providing QoS to a network is by differentiating the type of traffic by class or flow and assigning priorities to each type. When the network becomes congested, the data packets that are marked as having higher priority will have higher likelihood of being serviced.

3.16. Upstream

Description of the direction in which a signal travels, in this case from the user to the network. Upstream traffic occurs when users are uploading something to the network, such as sending email, sending files to another computer, or uploading photos to a digital photo website.

4. Historical Overview

Comcast began the engineering project to develop a new congestion management system in March 2008, the same month that Comcast hosted the 71st meeting of the IETF in Philadelphia, PA, USA. On May 28, 2008, Comcast participated in an IETF Peer-to-Peer Infrastructure Workshop [RFC5594], hosted by the Massachusetts Institute of Technology (MIT) in Cambridge, MA, USA.

In order to participate in this workshop, interested attendees were asked to submit a paper to a technical review team, which Comcast did on May 9, 2008, in [COMCAST_P2PI_PAPER]. Comcast subsequently attended and participated in this valuable workshop. During the workshop, Comcast outlined the high-level design for a new congestion management system [COMCAST_P2PI_PRES] and solicited comments and other feedback from attendees and other members of the Internet community (presentations were also posted to the IETF's P2Pi mailing list). The congestion management system outlined in that May 2008 workshop was later tested in trial markets and is in essence what was then deployed by Comcast later in 2008.

Following an August 2008 FCC document [FCC_Memo_Opinion] regarding how Comcast managed congestion on its High-Speed Internet ("HSI") network, Comcast disclosed to the FCC [FCC_Net_Mgmt_Response] and the public additional technical details of the congestion management system that it intended to and did implement by the end of 2008 [FCC_Congest_Mgmt_Ltr], including the thresholds involved in this new

system. While the description of how this system is deployed in the Comcast network is necessarily specific to the various technologies and designs specific to that network, a similar system could be deployed on virtually any large-scale ISP network or other IP network.

5. Summary

Comcast's HSI network has elements that are shared across many subscribers. This means that Comcast's HSI customers share upstream and downstream bandwidth with their neighbors. Although the available bandwidth is substantial, so, too, is the demand. Thus, when a relatively small number of customers in a neighborhood place disproportionate demands on network resources, this can cause congestion that degrades their neighbors' Internet experience. The goal of Comcast's new congestion management system is to enable all users of our network resources to access a "fair share" of that bandwidth, in the interest of ensuring a high-quality online experience for all of Comcast's HSI customers.

Importantly, the new approach is protocol-agnostic; that is, it does not manage congestion by focusing on the use of the specific protocols that place a disproportionate burden on network resources, or any other protocols. Rather, the new approach focuses on managing the traffic of those individuals who are using the most bandwidth at times when network congestion threatens to degrade subscribers' broadband experience and who are contributing disproportionately to such congestion at those points in time.

Specific details about these practices, including relevant threshold information, the type of equipment used, and other particulars, are discussed at some length later in this document. At the outset, however, we present a very high-level, simplified overview of how these practices work. Despite all the detail provided further below, the fundamentals of this approach can be summarized succinctly:

1. Software installed in the Comcast network continuously examines aggregate traffic usage data for individual segments of Comcast's HSI network. If overall upstream or downstream usage on a particular segment of Comcast's HSI network reaches a pre-determined level, the software moves on to step two.

2. At step two, the software examines bandwidth usage data for subscribers in the affected network segment to determine which subscribers are using a disproportionate share of the bandwidth.

If the software determines that a particular subscriber or subscribers have been the source of high volumes of network traffic during a recent period of minutes, traffic originating from that subscriber or those subscribers temporarily will be assigned a lower priority status.

3. During the time that a subscriber's traffic is assigned the lower priority status, their packets will not be delayed or dropped so long as the network segment is not actually congested. If, however, the network segment becomes congested, their packets could be intermittently delayed or dropped.

4. The subscriber's traffic returns to normal priority status once his or her bandwidth usage drops below a set threshold over a particular time interval.

Comcast undertook considerable effort, over the course of many months, to formulate our plans for this congestion management approach, adjusting them, and subjecting them to real-world trials. Market trials were conducted in Chambersburg, PA; Warrenton, VA; Lake City, FL; East Orange, FL; and Colorado Springs, CO, between June and September 2008. This enabled us to validate the utility of the general approach and collect substantial trial data to test multiple variations and alternative formulations.

6. Relationship between Managing Congestion and Adding Capacity

Many people have questioned whether congestion should ever exist at all, if an ISP was adding sufficient capacity. There is certainly a relationship between capacity and congestion. But there are two types of congestion that generally present themselves in a network.

The first general type of congestion is regularly occurring and is the result of gradually increasing traffic levels up to a point where typical usage peaks cause congestion on a regular basis. Comcast, like many ISPs, has a set capacity management process by which capacity additions are automatically triggered based on certain usage trends; this process is geared towards bringing additional capacity to the network prior to the onset of regularly occurring congestion. As such, capacity is added when needed and before it presents noticeable effects. This process is in place since capacity additions are not instantaneous and in many cases require significant physical work.

The second general type of congestion is unpredictable congestion, which can occur for a wide range of reasons. One example may be due to current events, where users may be all rushing to access specific content at the exact same time, and where the systems serving that

content may not be able to keep up with demand. Another example may be due to a localized disaster, where some network paths have been destroyed or otherwise impaired, and where many users are attempting to communicate with one another at traffic levels significantly above normal.

Thus, in both cases, even with continuous upgrades and constant investment in additional capacity, the fact remains that network capacity is not unlimited. A congestion management system, absent superior protocol-based solutions that do not currently exist, can therefore help manage the effects of congestion on users, improving their Internet experience.

7. Implementation and Configuration

It is important to note that the implementation details below and the overall design of the system are matched to traffic patterns that exist on the Internet today and that the authors believe will exist in the near future. While the authors desired to make the system highly adaptable and a good long-term network investment, significant changes in such traffic patterns may necessitate a change in the configuration of the system or, in extreme cases, a different type of system altogether.

To understand exactly how these new congestion management practices work, it is helpful to have a general understanding of how Comcast's HSI network is designed. Comcast's HSI network is what is commonly referred to as a hybrid fiber-coax network, with coaxial cable connecting each subscriber's cable modem to an Optical Node, and fiber-optic cables connecting the Optical Node, through distribution hubs, to the Cable Modem Termination System (CMTS), which is also known as a "data node". The CMTSs are then connected to higher-level routers, which in turn are connected to Comcast's Internet backbone facilities. Today, Comcast has over 3,200 CMTSs deployed throughout our network, serving over 15 million HSI subscribers.

Each CMTS has multiple "ports" that handle traffic coming into and leaving the CMTS. In particular, each cable modem deployed on the Comcast HSI network is connected to the CMTS through the ports on the CMTS. These ports can be either "downstream" ports or "upstream" ports, depending on whether they send information to cable modems (downstream) or receive information from cable modems (upstream) attached to the port. (Note that the term "port" as used here generally contemplates single channels on a CMTS, but these statements will apply to virtual channels, also known as "bonded

groups", in a DOCSIS 3.0 environment.) Even without channel bonding, multiple channels are usually configured to come out of each physical port. Said another way, there is generally a mapping of multiple channels to each physical port.

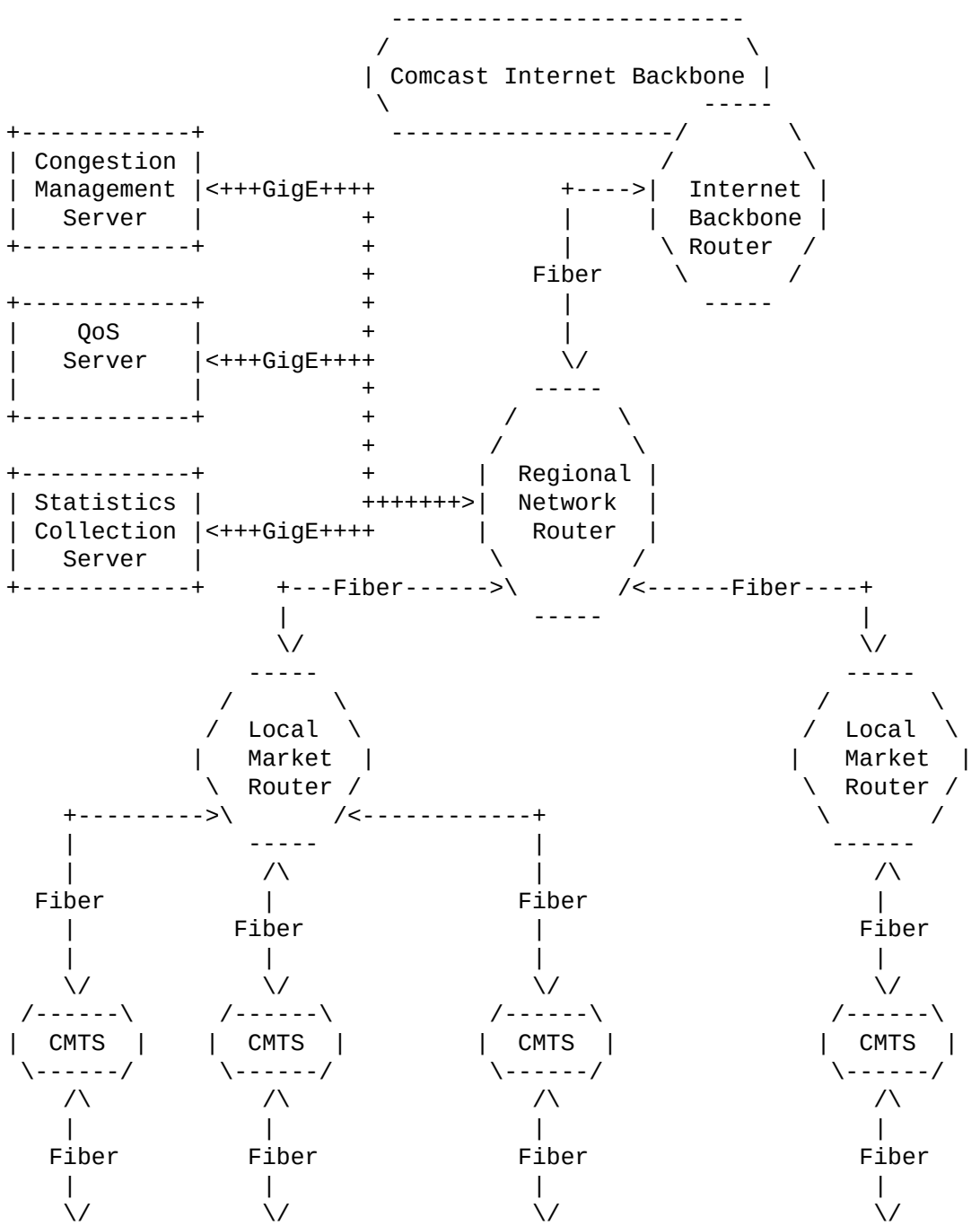
Currently, on average, approximately 275 cable modems share the same downstream port, and about 100 cable modems share the same upstream port; however, this is constantly changing (both numbers generally become smaller over time, based on current DOCSIS technology). Both types of ports can experience congestion that could degrade the broadband experience of our subscribers and, unlike with the previous congestion management practices, both upstream and downstream traffic are subject to management in this new congestion management system.

Based upon the design of the network and traffic patterns observed, the most likely place for congestion to occur is on these CMTS ports. As a result, the congestion management system measures the traffic conditions of CMTS ports, and applies any policy actions to traffic on those ports (rather than some other, more distant segment of the network).

To implement Comcast's new protocol-agnostic congestion management practices, Comcast purchased new hardware and software that were deployed near the Regional Network Routers ("RNRs") that are further upstream in Comcast's network. This new hardware consists of Internet Protocol Detail Record ("IPDR") servers, Congestion Management servers, and PacketCable Multimedia ("PCMM") servers. Further details about each of these pieces of equipment can be found below, in Section 7.4. It is important to note here, however, that even though the physical location of these servers is at the RNR, the servers communicate with -- and manage individually -- multiple ports on multiple CMTSs to effectuate the practices described in this document. That is to say, bandwidth usage on one CMTS port will have no effect on whether the congestion management practices described herein are applied to a subscriber on a different CMTS port.

Figure 1 provides a simplified graphical depiction of the network architecture just described:

Figure 1: Simplified Network Diagram Showing High-Level Comcast Network and Servers Relevant to Congestion Management



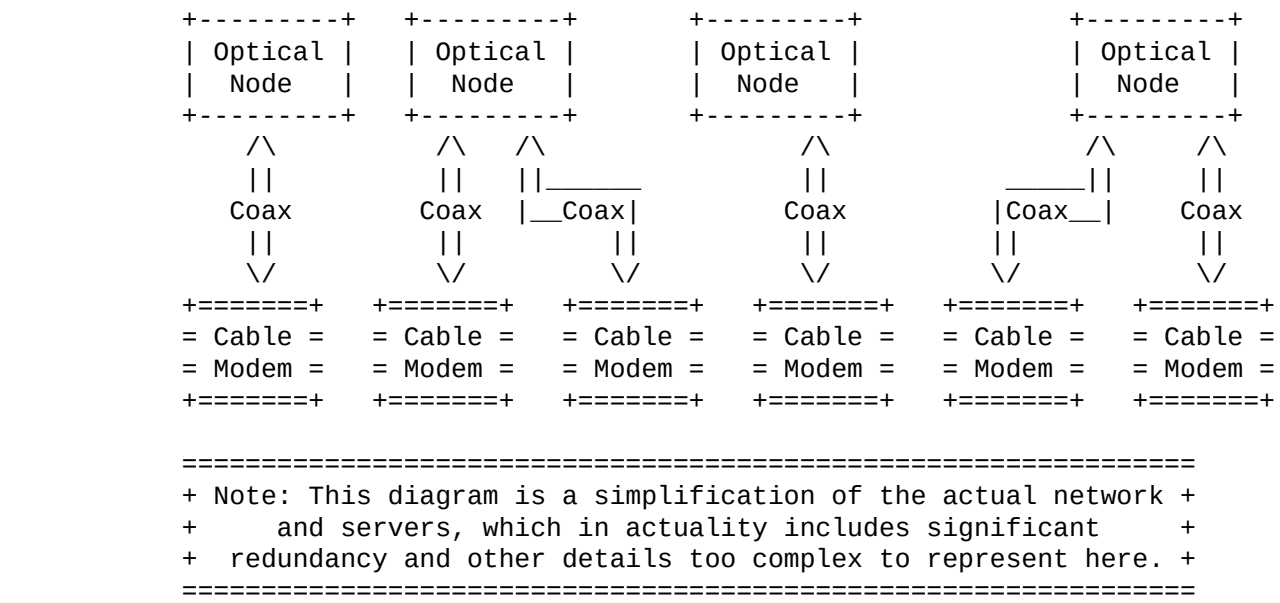


Figure 1

Each Comcast HSI subscriber's cable modem has a "bootfile", which is essentially a configuration file that contains certain pieces of information about the subscriber's service to ensure that the service functions properly. (Note: No personal information is included in the bootfile; it only includes information about the service that the subscriber has purchased.) For example, the bootfile contains information about the maximum speed (what we refer to in this document as the "provisioned bandwidth") that a particular modem can achieve based on the tier (personal/residential, commercial, etc.) the customer has purchased. Bootfiles are generally reset from time to time to account for changes in the network and other updates, and this is usually done through a command sent from the network and without the subscriber noticing. In preparation for the transition to this new congestion management system, Comcast sent new bootfiles to our HSI customers' cable modems that created two Quality of Service (QoS) levels for Internet traffic going to and from the cable modem: (1) "Priority Best Effort" ("PBE") traffic; and (2) "Best Effort" ("BE") traffic. As with previous changes to cable modem bootfiles, the replacement of the old bootfile with the new bootfile requires no active participation by Comcast customers.

Thereafter, all traffic going to or coming from cable modems on the Comcast HSI network is designated as either PBE or BE. PBE is the default status for all Internet traffic coming from or going to a particular cable modem. Traffic is designated BE for a particular cable modem only when both of two conditions are met:

- First, the usage level of a particular upstream or downstream port of a CMTS, as measured over a particular period of time, must be nearing the point where congestion could degrade users' experience. We refer to this as the "Near Congestion State" and, based on the technical trials we have conducted (further validated in our full deployment), we have established a threshold, described in more detail below, for when a particular CMTS port enters that state.

- Second, a particular subscriber must be making an extended, high contribution to the bandwidth usage on the particular port, relative to the service tier they purchased, as measured over a particular period of time. We refer to this as the "Extended High Consumption State" and, based on the technical trials we have conducted (further validated in our full deployment), we have established a threshold, described in more detail below, for when a particular user enters that state.

When, and only when, both conditions are met, a user's upstream or downstream traffic (depending on which type of port is in the Near Congestion State) is designated as BE. Then, to the extent that actual congestion occurs, any delay resulting from the congestion will affect BE traffic before it affects PBE traffic.

We now explain the foregoing in greater detail in the following sections.

7.1. Thresholds for Determining When a CMTS Port Is in a Near Congestion State

For a CMTS port to enter the Near Congestion State, traffic flowing to or from that CMTS port must exceed a specified level (the "Port Utilization Threshold") for a specific period of time (the "Port Utilization Duration"). The Port Utilization Threshold on a CMTS port is measured as a percentage of the total aggregate upstream or downstream bandwidth for the particular port during the relevant timeframe. The Port Utilization Duration on the CMTS is measured in minutes.

Values for each of the thresholds that are used as part of this congestion management technique have been tentatively established after an extensive process of lab tests, simulations, technical trials, vendor evaluations, customer feedback, and a third-party consulting analysis. In the same way that specific anti-spam or other network management practices are adjusted to address new issues that arise, it is a near certainty that these values will change over time, as Comcast gathers more data and performs additional analysis resulting from wide-scale use of the new technique. Moreover, as

with any large network or software system, software bugs and/or unexpected errors may arise, requiring software patches or other corrective actions. As always, Comcast's decisions on these matters are driven by the marketplace imperative that we deliver the best possible experience to our HSI subscribers.

Given our experience as described above, we determined that a starting point for the upstream Port Utilization Threshold should be 70 percent and the downstream Port Utilization Threshold should be 80 percent. For the Port Utilization Duration, we determined that the starting point should be approximately 15 minutes (although some technical limitations in some newer CMTSs deployed on Comcast's network may make this time period vary slightly). Thus, over any 15-minute period, if an average of more than 70 percent of a port's upstream bandwidth capacity or more than 80 percent of a port's downstream bandwidth capacity is utilized, that port is determined to be in a Near Congestion State.

Based on the trials conducted and operational experience to date, a typical CMTS port on our HSI network is in a Near Congestion State only for relatively small portions of the day, if at all, though there is no way to forecast what will be the busiest time on a particular port on a particular day. Moreover, the trial data and operational experience indicate that, even when a particular port is in a Near Congestion State, the instances where the network actually becomes congested during the Port Utilization Duration are few, and managed users whose packets may be intermittently delayed or dropped during those congested periods perceive little, if any, effect, as discussed below.

7.2. Thresholds for Determining When a User Is in an Extended High Consumption State and for Release from That Classification

Once a particular CMTS port is in a Near Congestion State, the software examines whether any cable modems are consuming bandwidth disproportionately. (Note: Although each cable modem is typically assigned to a particular household, the software does not and cannot actually identify individual users or the number of users sharing a cable modem, or analyze particular users' traffic.) For purposes of this document, we use "cable modem", "user", and "subscriber" interchangeably to mean a subscriber account or user account and not an individual person. For a user to enter an Extended High Consumption State, he or she must consume greater than a certain percentage of his or her provisioned upstream or downstream bandwidth (the "User Consumption Threshold") for a specific length of time (the "User Consumption Duration"). The User Consumption Threshold is measured as a user's consumption of a particular percentage of his or her total provisioned upstream or downstream bandwidth. That

bandwidth is the maximum speed that a particular modem can achieve based on the tier (personal/residential, commercial, etc.) the customer has purchased. For example, if a user buys a service with speeds of 50 Mbps downstream and 10 Mbps upstream, then his or her provisioned downstream speed is 50 Mbps and provisioned upstream speed is 10 Mbps. It is also important to note that because the User Consumption Threshold is a percentage of provisioned bandwidth for a particular user account, and not a static value, users of higher-speed tiers have correspondingly higher User Consumption Thresholds. Lastly, the User Consumption Duration is measured in minutes.

Following lab tests, simulations, technical trials, customer feedback, vendor evaluations, and an independent third-party consulting analysis, we have determined that the appropriate starting point for the User Consumption Threshold is 70 percent of a subscriber's provisioned upstream or downstream bandwidth, and that the appropriate starting point for the User Consumption Duration is 15 minutes (this has been further validated in our full deployment). That is, when a subscriber uses an average of 70 percent or more of his or her provisioned upstream or downstream bandwidth over a particular 15-minute period, that user is then in an Extended High Consumption State. Therefore, this is a consumption-based threshold and not a peak-speed-based threshold. Thus, the Extended High Consumption State is not tied to whether a user has bursted once or more above this 70% threshold for a brief moment. Instead, it is consumption-based, meaning that a certain bitrate must be exceeded over at least the entire User Consumption Duration.

The User Consumption Thresholds have been set sufficiently high that using the HSI connection for Voice over IP (VoIP), gaming, web surfing, or most streaming video cannot alone cause subscribers to our standard-level HSI service to exceed the User Consumption Threshold. For example, while one of Comcast's common HSI service tiers has a provisioned downstream bandwidth of 22 Mbps today, streaming video (even some HD video) from Hulu uses less than 2.5 Mbps, a Vonage or Skype VoIP call uses less than 131 kbps, and streaming music uses less than 128 kbps (in this example, 70 percent of 22 Mbps is 15.4 Mbps). As noted above, these values are subject to change as necessary in the same way that specific anti-spam or other network management practices are adjusted to address new issues that arise, or should unexpected software bugs or other problems arise.

Based on data collected from the trial markets where the new congestion management practices were tested (further validated in our full deployment), on average less than one-third of one percent of subscribers have had their traffic priority status changed to the BE state on any given day. For example, in Colorado Springs, CO, the

largest test market, on any given day in August 2008, an average of 22 users out of 6,016 total subscribers in the trial had their traffic priority status changed to BE at some point during the day.

A user's traffic is released from a BE state when the user's bandwidth consumption drops below 50 percent of his or her provisioned upstream or downstream bandwidth for a period of approximately 15 minutes. These release criteria are intended to minimize (and hopefully prevent) user QoS oscillation, i.e., a situation in which a particular user could cycle repeatedly between BE and PBE. Thus, without this lower release criteria, we were concerned that certain users would oscillate between BE and PBE states for an extended period, without clear benefit to the system and other users, and would place an unnecessary signaling burden on the system. NetForecast, Inc., an independent consultant retained to provide analysis and recommendations regarding Comcast's trials and related congestion management work, suggested this approach, which has worked well in our trials, lab testing, and subsequent national deployment.

Simply put, there are four steps for determining whether the traffic associated with a particular cable modem is designated as PBE or BE:

1. Determine if the CMTS port is in a Near Congestion State.

2. If yes, determine whether any users are in an Extended High Consumption State.

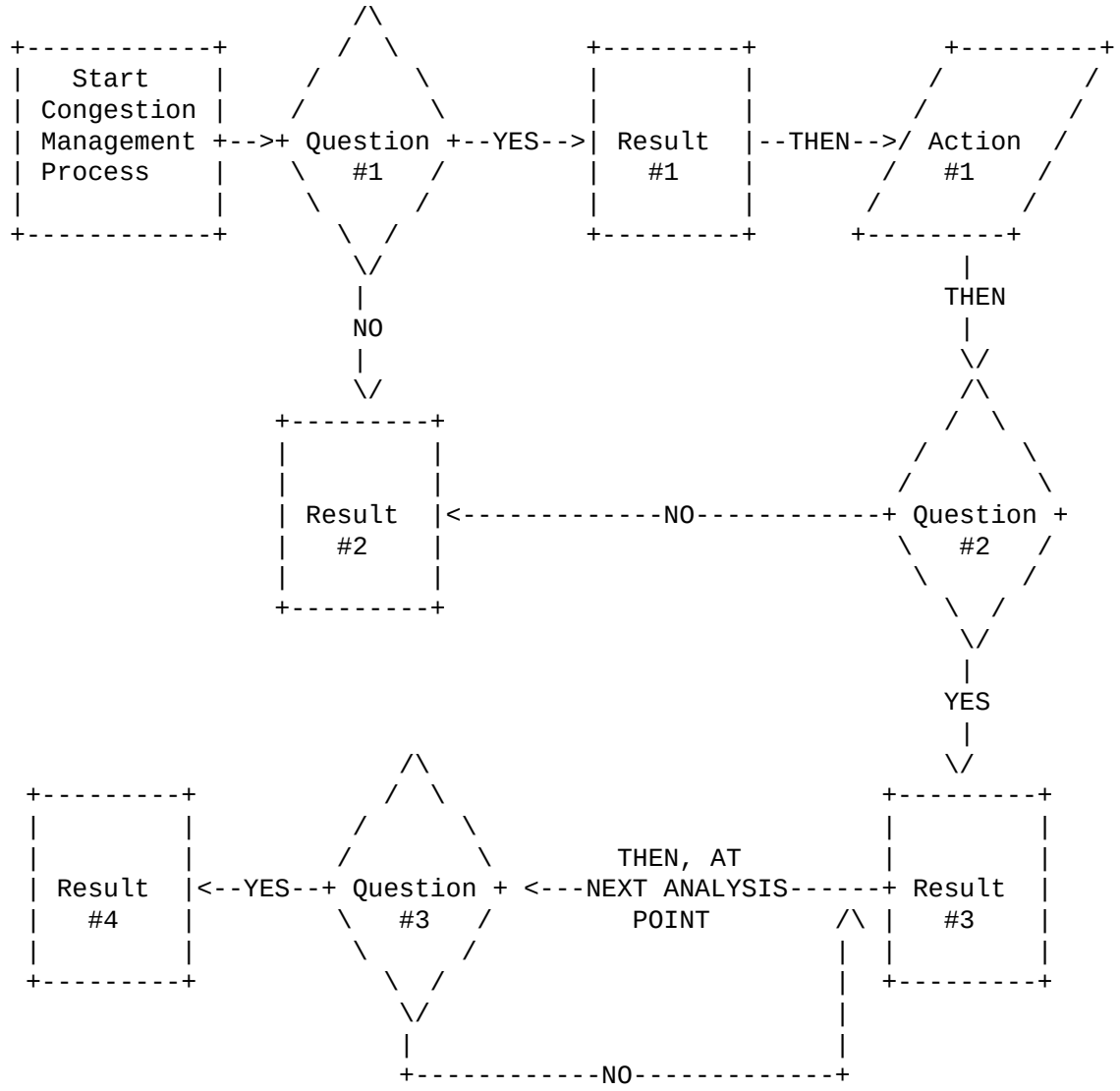
3. If yes, change those users' traffic to BE from PBE. If the answer at either step one or step two is no, no action is taken.

4. If a user's traffic has been designated BE, check user consumption at the next interval. If user consumption has declined below the predetermined threshold, reassign the user's traffic as PBE. If not, recheck at the next interval.

In cases where a CMTS regularly enters a Near Congestion State, and where congestion subsequently does occur, but where no users match the criteria to be classified in an Extended High Consumption State, this may indicate the congestion observed is regularly occurring, rather than unpredictable congestion. As such, this may be an additional data point in favor of considering whether and when to add capacity.

Figure 2 graphically depicts how this congestion management process works, using an example of a situation where upstream port utilization may be reaching a Near Congestion State (the same diagram, with different values in the appropriate places, could be used to depict the management process for downstream ports, as well):

Figure 2: Upstream Congestion Management Decision Flowchart



KEY TO FIGURE 2 ABOVE:

Question #1: Is the CMTS Upstream Port Utilization at an average of OVER 70% for OVER 15 minutes?

Result #1: CMTS marked in a Near Congestion State, indicating congestion *may* occur soon.

Action #1: Search most recent analysis timeframe (approx. 15 mins.) of IPDR usage data.

Question #2: Are any users consuming an average of OVER 70% of provisioned upstream bandwidth for OVER 15 minutes?

Result #2: No action taken.

Result #3: Change user's upstream traffic from Priority Best Effort (PBE) to Best Effort (BE).

Question #3: Is the user in Best Effort (BE) consuming an average of LESS THAN 50% of provisioned upstream bandwidth over a period of 15 minutes?

Result #4: Change user's upstream traffic back to Priority Best Effort (PBE) from Best Effort (BE).

Figure 2

7.3. Effect of BE Quality of Service on Users' Broadband Experience

When a CMTS port is in a Near Congestion State and a cable modem connected to that port is in an Extended High Consumption State, that cable modem's traffic is designated as BE. Depending upon the level of utilization on the CMTS port, this designation may or may not result in the user's traffic being delayed or, in extreme cases, dropped before PBE traffic is dropped. This is because of the way that the CMTS handles traffic. Specifically, CMTS ports have what is commonly called a "scheduler" that puts all the packets coming from or going to cable modems on that particular port in a queue and then handles them in turn. A certain number of packets can be processed by the scheduler in any given moment; for each time slot, PBE traffic is given priority access to the available capacity, and BE traffic is processed on a space-available basis.

A rough analogy would be to busses that empty and fill up at incredibly fast speeds. As empty busses arrive at the figurative "bus stop" -- every two milliseconds in this case -- they fill up with as many packets as are waiting for "seats" on the bus, to the

limits of the bus' capacity. During non-congested periods, the bus will usually have several empty seats, but during congested periods, the bus will fill up and packets will have to wait for the next bus. It is during the congested periods that BE packets will be affected. If there is no congestion, packets from a user in a BE state should have little trouble getting on the bus when they arrive at the bus stop. If, on the other hand, there is congestion in a particular instance, the bus may become filled by packets in a PBE state before any BE packets can get on. In that situation, the BE packets would have to wait for the next bus that is not filled by PBE packets. In reality, this all takes place in two-millisecond increments, so even if the packets miss 50 "busses", the delay will only be about one-tenth of a second.

During times of actual network congestion, when packets from BE traffic might be intermittently delayed, there is a variety of effects that could be experienced by a user whose traffic is delayed, depending upon what applications he or she is using. Typically, a user whose traffic is in a BE state during actual congestion may find that a webpage loads sluggishly, a peer-to-peer upload takes somewhat longer to complete, or a VoIP call sounds choppy. Of course, the same thing could happen to the customers on a port that is congested in the absence of any congestion management; the difference here is that the effects of any such delays are shifted toward those who have been placing the greatest burden on the network, instead of being distributed randomly among the users of that port without regard to their consumption levels. As a matter of fact, our studies concluded that the experience of the PBE subscribers improves when this congestion management system is enabled. This conclusion is based on network measurements, such as latency.

NetForecast explored the potential risk of a worst-case scenario for users whose traffic is in a BE state: the possibility of "bandwidth starvation" in the theoretical case where 100 percent of the CMTS bandwidth is taken up by PBE traffic for an extended period of time. In theory, such a condition could mean that a given user whose traffic is designated BE would be unable to effectuate an upload or download (as noted above, both are managed separately) for some period of time. However, when these management techniques were tested, first in company testbeds and then in our real-world trials conducted in the five markets (further validated in our full deployment), such a theoretical condition did not occur. In addition, our experience with the system as fully deployed in our production network demonstrates that these management practices have very modest real-world impacts. In addition, Comcast did not receive a single customer complaint, in any of the trial markets, that could be traced to this congestion management system, despite having broadly publicized these trials. In our subsequent national

deployment into our production network, we still have yet to find a specific complaint that can be traced back to the effect of this congestion management system.

Comcast continues to monitor how user traffic is affected by these new congestion management techniques and will make the adjustments necessary to ensure that all Comcast HSI customers have a high-quality Internet experience.

7.4. Equipment/Software Used and Location

The above-mentioned functions are carried out using three different types of application servers, supplied by three different vendors. As mentioned above, these servers are installed near Comcast's regional network routers. The exact locations of these servers are not particularly relevant to this document, as this information does not change the fact that the servers manage individual CMTS ports.

The first application server is an IPDR server, which collects relevant cable modem volume usage information from the CMTS, such as how many aggregate upstream or downstream bytes a subscriber uses over a particular period of time. IPDR has been adopted as a standard by many industry organizations and initiatives, such as CableLabs, the Alliance for Telecommunications Industry Solutions (ATIS), the International Telecommunication Union (ITU), and the Third Generation Partnership Project (3GPP), among others. The IPDR software deployed was developed by Active Broadband Networks, and is noted as the Statistics Collection Server in Figure 3.

The second application server is the Congestion Management server, which uses the Simple Network Management Protocol (SNMP) [RFC3410] to measure CMTS port utilization and detect when a port is in a Near Congestion State. When this happens, the Congestion Management server then queries the relevant IPDR data for a list of cable modems meeting the criteria set forth above for being in an Extended High Consumption State. The Congestion Management server software deployed was developed by Sandvine.

If one or more users meet the criteria to be managed, then the Congestion Management server notifies a third application server, the PCMM application server, as to which users have been in an Extended High Consumption State and whose traffic should be treated as BE. The PCMM servers are responsible for signaling a given CMTS to set the traffic for specific cable modems with a BE QoS, and for tracking and managing the state of such CMTS actions. If no users meet the criteria to be managed, no users will have their traffic managed. The PCMM software deployed was developed by Camiant, and is noted as the QoS Server in Figure 3.

Figure 3 graphically depicts the high-level management flows among the congestion management components on Comcast's network, as described above:

Figure 3: Simplified Diagram Showing High-Level Management Flows Relevant to the System

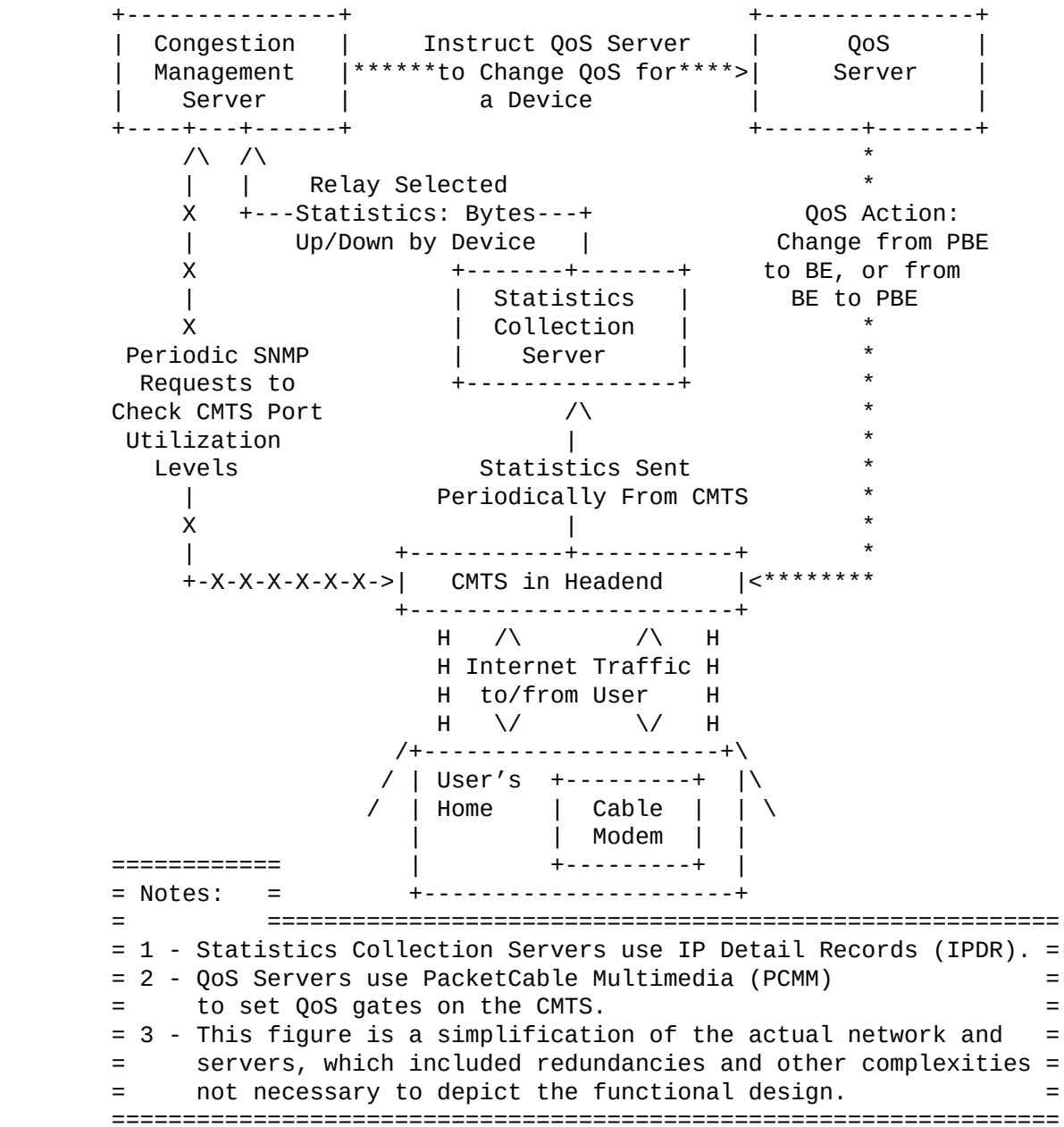


Figure 3

8. Conclusion

Comcast started design and development of this new protocol-agnostic congestion management system in March 2008. Comcast shared the design with the IETF and others in the Internet community, as well as with an independent consultant, incorporating feedback we received into the final design. Following lab testing, the system was tested in Comcast's production network in trial markets between June and September 2008. Comcast's production network transition to this new protocol-agnostic congestion management system began in October 2008 and was completed on December 31, 2008.

As described herein, the new approach does not manage congestion by focusing on managing the use of specific protocols. Nor does this approach use TCP "reset packets" [RFC3360]. Rather, the system acts such that during periods when a CMTS port is in a Near Congestion State, the system (1) identifies the subscribers on that port who have consumed a disproportionate amount of bandwidth over the preceding 15 minutes and (2) lowers the priority status of those subscribers' traffic to BE status until those subscribers meet the release criteria. During periods of actual congestion, the system handles PBE traffic before BE traffic. Comcast's trials and subsequent national deployment indicate that this new congestion management system ensures a quality online experience for all of Comcast's HSI customers.

9. Exceptional Network Utilization Considerations

This system was developed to cope with somewhat "normal" occurrences of congestion that could occur on virtually any IP network. It should also be noted, however, that such a system could also prove particularly useful in the case of "exceptional network utilization" events that existing network usage models do not or cannot accurately predict. Some network operators refer to these exceptional events as "surges" in utilization, similar to sudden surges in demand in electrical power grids, with which many people may be familiar.

For example, in the case of a severe global pandemic, it may be expected that large swaths of the population may need to work remotely, via their Internet connection. In such a case, a largely unprecedented level of utilization may occur. In such cases, it may be helpful to have a flexible congestion management system that could adapt to this situation and help allocate network resources while additional capacity is being brought online or while a temporary condition persists.

10. Limitations of This Congestion Management System

The main limitations of the system include:

- The system is not an end-to-end congestion management system, nor does it enable one.
- The system does not signal the presence of congestion to user applications or to all devices on the network path.
- The system does not explicitly enable additional user and/or application responses to congestion.
- The system does not enable distributed denial-of-service (DDoS) mitigation or other capabilities.

11. Low Extra Delay Background Transport and Other Possibilities

There are several new IETF working group efforts that are focused on the question of congestion and its effects, avoiding congestion, managing congestion, and communicating congestion information. This includes the Congestion Exposure (CONEX) working group, the Application Layer Transport Optimization (ALTO) working group, and the Low Extra Delay Background Transport (LEDBAT) working group. Should one or more of these working groups be successful in producing useful work, it is possible that the design or configuration of the system documented here may need to change. For example, this congestion management system does not currently have a way to take into account differing classes of data transfer, such as a class of data transfer that LEDBAT may specify, which may better yield to other traffic than existing transport protocols. In addition, CONEX may specify methods for this or other systems to signal congestion state or expected congestion to other parts of the network, and/or to hosts on either end of a particular network flow. Furthermore, it is conceivable that the result of current or future IETF work could obviate the need for such a congestion management system entirely.

12. Security Considerations

It is important that an ISP secure access to the Congestion Management servers and the QoS Servers, as well as QoS signaling to the CMTSs, so that unauthorized users and/or hosts cannot make unauthorized changes to QoS settings in the network.

It is also important to secure access to the Statistics Collection Server since this contains IPDR-based byte transfer data that is considered private by end users on an individual basis. In addition, this data is considered ISP-proprietary traffic data on an aggregate

basis. Access to the Statistics Collection Server should also be secured so that false usage statistics cannot be fed into the system. It is important to note that IPDR data contains a count of bytes sent and bytes received, by cable modem MAC address, over a given interval of time. This data does not contain things such as the source and/or destination Internet address of that data, nor does it contain the protocols used, ports used, etc.

13. Acknowledgements

The authors wish to acknowledge the hard work of the many people who helped to develop and/or review this document, as well as the people who helped deploy the system in such a short period of time.

The authors also wish to acknowledge the following individuals for performing a detailed review of this document and/or providing comments and feedback that helped to improve and evolve this document:

- Kris Bransom

- Bob Briscoe

- Lars Eggert

- Ari Keranen

- Tero Kivinen

- Matt Mathis

- Stanislav Shalunov

14. Informative References

[COMCAST_P2PI_PAPER]
Livingood, J. and R. Woundy, "Comcast's IETF P2P Infrastructure Workshop Position Paper", FCC Filings Comcast Network Management Proceedings, May 2008, <http://trac.tools.ietf.org/area/rai/trac/raw-attachment/wiki/PeerToPeerInfrastructure/16%20ietf-p2pi-comcast-20080509.pdf>.

[COMCAST_P2PI_PRES]
Livingood, J. and R. Woundy, "Comcast's IETF P2P Infrastructure Workshop Presentation on May 28, 2008", FCC Filings Comcast Network Management Proceedings, May 2008, <http://trac.tools.ietf.org/area/rai/trac/raw-attachment/wiki/PeerToPeerInfrastructure/02-Comcast-IETF-P2Pi.pdf>.

[DOCSIS_CM2CPE]
CableLabs, "Data-Over-Cable Service Interface Specifications - DOCSIS 3.0 - Cable Modem to Customer Premise Equipment Interface Specification", DOCSIS 3.0 CM-SP-CMCIv3-I01-080320, March 2008, <http://www.cablelabs.com/cablemodem/specifications/specifications30.html>.

[DOCSIS_IPDR]
Yassini, R., "Data-Over-Cable Service Interface Specifications - DOCSIS 2.0 - Operations Support System Interface Specification", DOCSIS 2.0 CM-SP-OSSIv2.0-C01-081104, November 2008, <http://www.cablelabs.com/cablemodem/specifications/specifications30.html>.

[DOCSIS_MULPI]
CableLabs, "Data-Over-Cable Service Interface Specifications - DOCSIS 3.0 - MAC and Upper Layer Protocols Interface Specification", DOCSIS 3.0 CM-SP-MULPIv3.0-I11-091002, October 2009, <http://www.cablelabs.com/cablemodem/specifications/specifications30.html>.

[DOCSIS_OSSI]
CableLabs, "Data-Over-Cable Service Interface Specifications - DOCSIS 3.0 - Operations Support System Interface Specification", DOCSIS 3.0 CM-SP-OSSIv3.0-I10-091002, October 2009, <http://www.cablelabs.com/cablemodem/specifications/specifications30.html>.

[DOCSIS_PHY]
CableLabs, "Data-Over-Cable Service Interface Specifications - DOCSIS 3.0 - Physical Layer Specification", DOCSIS 3.0 CM-SP-PHYv3.0-I08-090121, January 2009, <http://www.cablelabs.com/cablemodem/specifications/specifications30.html>.

[DOCSIS_SEC]
CableLabs, "Data-Over-Cable Service Interface Specifications - DOCSIS 3.0 - Security Specification", DOCSIS 3.0 CM-SP-SECv3.0-I11-091002, March 2008, <http://www.cablelabs.com/cablemodem/specifications/specifications30.html>.

[FCC_Congest_Mgmt_Ltr]
Zachem, K., "Letter to the FCC Advising of Successful Deployment of Comcast's New Congestion Management System", FCC Filings Comcast Network Management Proceedings, January 2009, <http://fjallfoss.fcc.gov/ecfs/document/view?id=6520192582>.

[FCC_Memo_Opinion]
Martin, K., Copps, M., Adelstein, J., Tate, D., and R. McDowell, "FCC Memorandum and Opinion Regarding Reasonable Network Management", File No. EB-08-IH-1518 WC Docket No. 07-52, August 2008, <http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-08-183A1.pdf>.

[FCC_Net_Mgmt_Response]
Zachem, K., "Letter to the FCC Regarding Comcast's Network Management Practices", FCC Filings Comcast Network Management Proceedings, September 2008, <http://fjallfoss.fcc.gov/ecfs/document/view?id=6520169715>.

[IPDR_Standard]
Cotton, S., Cockrell, B., Walls, P., and T. Givoly, "Network Data Management - Usage (NDM-U) For IP-Based Services. Service Specification - Cable Labs DOCSIS 2.0 SAMIS", IPDR Service Specifications NDM-U, November 2004, <http://www.ipdr.org/public/Service_Specifications/3.X/DOCSIS(R)3.5-A.0.pdf>.

[PowerBoost_Specification]
Comcast Cable Communications Management LLC, "Comcast PowerBoost Specification", Website Comcast.com, June 2010, <http://customer.comcast.com/Pages/FAQListViewer.aspx?topic=Internet&folder=8b2fc392-4cde-4750-ba34-051cd5feacf0>.

[RFC1633] Braden, B., Clark, D., and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC3083] Woundy, R., "Baseline Privacy Interface Management Information Base for DOCSIS Compliant Cable Modems and Cable Modem Termination Systems", RFC 3083, March 2001.

[RFC3360] Floyd, S., "Inappropriate TCP Resets Considered Harmful", BCP 60, RFC 3360, August 2002.

[RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart, "Introduction and Applicability Statements for Internet-Standard Management Framework", RFC 3410, December 2002.

[RFC5594] Peterson, J. and A. Cooper, "Report from the IETF Workshop on Peer-to-Peer (P2P) Infrastructure, May 28, 2008", RFC 5594, July 2009.

Authors' Addresses

Chris Bastian
Comcast Cable Communications
One Comcast Center
1701 John F. Kennedy Boulevard
Philadelphia, PA 19103
US
EMail: chris_bastian@cable.comcast.com
URI: http://www.comcast.com

Tom Klieber
Comcast Cable Communications
1306 Goshen Parkway
West Chester, PA 19380
US
EMail: tom_klieber@cable.comcast.com
URI: http://www.comcast.com

Jason Livingood
Comcast Cable Communications
One Comcast Center
1701 John F. Kennedy Boulevard
Philadelphia, PA 19103
US
EMail: jason_livingood@cable.comcast.com
URI: http://www.comcast.com

Jim Mills
Comcast Cable Communications
One Comcast Center
1800 Bishops Gate Drive
Mount Laurel, NJ 08054
US
EMail: jim_mills@cable.comcast.com
URI: http://www.comcast.com

Richard Woundy
Comcast Cable Communications
27 Industrial Avenue
Chelmsford, MA 01824
US
EMail: richard_woundy@cable.comcast.com
URI: http://www.comcast.com