

Sandvine Traffic Optimization



Contents

What is Network Neutrality?	1
Today's Traffic Trends	1
The Need for Traffic Optimization	2
Forms of Traffic Optimization	3
What is Policy Management?	4
Types of Policies	4
What about DOCSIS 3.0?	6
Optimizing Traffic while Maintaining Network Neutrality	6

The Value of Traffic Optimization in a World with Network Neutrality

The continued popularity of bandwidth intensive applications has contributed to the success of the Internet with unprecedented levels of traffic growth, innovative applications and growth in broadband subscribers. The evolving Internet environment unexpectedly changes usage patterns and causes network congestion that in turn affects the quality of experience for subscribers. Because all traffic is not created equal, service providers must find a balance between maximizing the value of the network for all involved and maintaining the network neutrality expected by its users.

The Value of Traffic Optimization in a World with Network Neutrality

Where there is traffic, inevitably there will be congestion. Nowhere is this truer than in today's broadband networks. The continued popularity of bandwidth-intensive applications has contributed to the success of the Internet with unprecedented levels of traffic growth, innovative applications and growth in broadband subscribers. The evolving Internet environment unexpectedly changes usage patterns and causes network congestion that in turn affects the quality of experience for subscribers. Because all traffic is not created equal, service providers must find a balance between maximizing the value of the network for all involved and maintaining the network neutrality expected by its users.

What is Network Neutrality?

To date, the FCC has avoided using regulation to manage the broadband industry, opting instead to adopt four principles of network neutrality in their policymaking. According to these four "Internet Freedoms", users are entitled to:

- (1) Access the lawful Internet content of their choice
- (2) Run applications and services of their choice, subject to the needs of law enforcement
- (3) Connect their choice of legal devices that do not harm the network
- (4) Competition among network providers, application and service providers, and content providers

These principles are intended both to encourage broadband deployment and to preserve and promote the open and interconnected nature of the public Internet.¹ But what happens to these principles when network congestion starts impacting overall user satisfaction, or when a minority of Internet users starts impacting the majority of Internet users? Service providers are now grappling with these questions as they come to terms with the effects of bandwidth intensive applications on their network traffic.

Today's Traffic Trends

The advent of bandwidth-intensive applications has radically altered bandwidth demand and traffic patterns for today's networks—directly impacting network infrastructure. DSL, cable and wireless networks are all hampered by a design philosophy that no longer reflects current bandwidth usage. The asymmetrical design of these networks, which dictates that downstream traffic is faster than upstream traffic, was originally based on usage patterns from early content-consuming applications like e-mail and Web-browsing. However, the continual evolution of applications from content consuming to always-on content supplying has meant that current traffic patterns no longer fit asymmetrical bandwidth assumptions. The result is that some network traffic gets delayed, dropped or degraded.

This situation is unlikely to improve. Bandwidth-intensive applications are appearing at a rate faster than capacity can be added to the last mile of networks. According to Sandvine's latest research, peer-to-peer (P2P) traffic continues to dominate upstream at more than 70% traffic. When combining upstream and downstream consumption rates for a broad indicator of network activity, P2P now accounts for more than 40% of the total bandwidth. With P2P traffic running 24x7, and its peak usage coinciding with the peak usage of other applications, there is a high probability that subscribers are going to experience congestion when they most want to use the Internet.

Compounding the congestion crisis is the longstanding policy of overbooking networks. Overbooking relies on the idea that there will never be enough subscribers online at any given time to use up all available network resources. Network providers are free to oversubscribe their networks until peak capacity is achieved. Once again, bandwidth-intensive applications like P2P are confounding these models by seriously impacting the sustainability of available bandwidth and eroding network performance and the overall user experience.²

It is clear that there is no escaping network congestion. But the fundamental question is not whether to manage the network but how to manage the network. The default is to do nothing and effectively apply a “first come first serve” (also known as, first-in first-out, FIFO) policy. Not surprisingly, FIFO policies encourage the greediest bandwidth applications to continue to use a disproportionate share of the bandwidth at the expense of other applications. The alternative is to implement some form of traffic management to ensure a more equitable distribution of network resources and maximize the value of the network for all involved.

The Need for Traffic Optimization

Broadband networks are not alone in their battle with congestion and overbooking—both are common to many industries. A classic example is the highway system. The assumption in highway design is that not everyone can drive on highways simultaneously. Even so, congestion and traffic jams persist. To mitigate congestion, highway designers incorporate mechanisms to control the traffic speed and volume on highways. These options vary depending on the severity of congestion and include everything from posted speed limits, to the timed entry of cars at on-ramps, to the closing of on-ramps when congestion is at its peak. The overall objective for imposing these measures is to ensure that everyone gets a fair chance at using a shared resource and can use it in a manner as designed.

Broadband networks share this same objective. But how this objective is to be achieved and by what mechanism has remained highly debated. One approach has been the concept of “flow-based fairness” in which relative flow rates are controlled in the attempt to attain fair resource allocation. This was the goal behind widely deployed protocols like weighted fair queuing (WFQ), TCP congestion control and TCP-friendly rate control.³

The fundamental flaw with this logic is that it assumes that subscribers use a low number of concurrent flows in each of the online applications they enjoy. Although many applications originally used a low number of concurrent flows, the introduction of protocols like Bit-Torrent, that open many TCP connections (or “flows”), simultaneously breaks this paradigm and invalidates the idea that user-to-user fairness can be met through flow-based fairness.

The most respected approach to-date for fair Internet access (specific to elastic demand) was pioneered by Frank Kelly⁴ and has been extended by others like Bob Briscoe⁵ at BT. It argues that welfare is optimized if each user is penalized (i.e. charged) for the value of the traffic that his/her traffic is denying. Service providers may not want to adopt this exact policy due to the user education and technology required to have it working to the desired level of precision.⁶ Today’s traffic optimization measures can replicate this idea and as network congestion remains inevitable, traffic optimization offers the best chance of reducing the impacts of congestion on subscribers and ensuring fair usage of a shared and limited resource.

Traffic optimization has a host of other benefits for both service providers and subscribers alike, including:

- **Maximizing the quality of experience for the users at large.** Traffic optimization ensures that all users will have a fair opportunity to use the shared service in a manner that maximizes their own utility. It is especially important during periods of peak usage

and under abnormal conditions such as public emergencies, link outages, etc. For example, one Bit Torrent session can impact 1000 phone calls.

- **Maximizing the efficiency of the network by optimizing the use of the bandwidth in the last mile.** Congestion typically occurs in the last mile of networks. Traffic optimization allows service providers to defer bandwidth expansion and capital expenditure until the marginal benefit for all the users exceeds the marginal cost of the bandwidth expansion. Service providers do not have to size their network to meet occasional peak busy hours and leave unused bandwidth in the network.
- **Providing a competitive advantage.** In optimizing traffic, service providers can differentiate their service from others by offering more than just speed, giving users a better overall quality of experience than they would otherwise receive from competing service providers.

Forms of Traffic Optimization

Service providers have at their disposal a number of different traffic optimization options. The most commonly used include quota/consumption caps, traffic prioritization, traffic policing, and traffic shaping.

- **Quota/Consumption caps** are metered usage for a billing period. Subscribers are allotted a quota of bytes or bandwidth they can use for the billing period. When the subscriber reaches the limit, they either have to buy more or stop using their service. Quota management by itself does not address network congestion, as it does not reduce the network bandwidth requirements during busy periods. Quota management only limits the overall amount of bandwidth a subscriber uses.
- **Traffic prioritization** is a method of applying different classes of services to packets and giving each class a different priority. The classification of packets can be done on a per-flow, application, device, and/or user basis. When there is congestion, the higher priority packets take priority over the lower priority packets. To ensure that the lower priority packets get some level of service during congestion and are not blocked or starved, traffic prioritization is commonly used in conjunction with some type of class-based queuing.
- **Traffic policing** is a method of ensuring that classified packets do not exceed a desired bit rate. Essentially, it is used to control the speed of traffic entering a network. Policing is currently accomplished using bit rate limits on all traffic coming to or from a subscriber. While this allows the highest granularity and the broadest controls for traffic optimization, if over-applied, traffic policing may not meet the network neutrality principles.
- **Traffic shaping** manages the transmission rate in order to optimize or guarantee performance. It is similar to traffic policing, but instead of dropping packets that exceed the bit rate limit the packets are queued and metered out so as not to exceed the bit rate limit.

Today's technology allows service providers to expand the attributes on which traffic optimization is applied to include the application, the time of day, the end device type and other traffic conditions. Shaping can also be further enhanced with the addition of prioritization algorithms, such as weighted fair queuing, that prioritize the traffic to be shaped. These enhancements are generally enforced through the use of policies.

What Is Policy Management?

One of the key tools used within the context of traffic optimization is policy management. Policy management allows high-level business requirements or rules (such as service level agreements) to be translated and enforced by the network. Service providers define a set of policies to determine how the network should operate under different conditions, such as traffic congestion, exceeded quotas, or security attacks. For a further layer of granularity, service providers can also choose to add subscriber service information to these policies.

Policies take the form of a set of conditions that trigger an action. The conditions include network events such as utilization, congestion, applications starting/stopping, users registering on the network, location on the network, time of day, service plan, and bill status. Policies are constructed to trigger actions to meet either the network optimization goals and/or the service guarantees for the subscribers.

Types of Policies

Policies are enforced using either an inline enforcement device or an offline or out-of-band enforcement device. Policies can be applied at the layer 3 or the IP-flow level in both the upstream and downstream direction, at the application level, or at the subscriber level. Commonly used policies include quota or consumption caps and Fairshare.

Offline versus Inline Policies

Policies performed by an offline device are commonly referred to as offline policies. Likewise, policies applied by an inline device are commonly referred to as inline policies.

Offline policies are performed by a device that has a “tap” on the data plane that allows it to view, but not manipulate the traffic. Policy management systems inspect and analyze this data plane for specific network conditions to trigger policy actions. The resulting policy actions can then either inject session management policies or can be used to trigger policy enforcement requests to other networks in the data plane such as the CMTS or to a quota management system.

Inline policies are enforced by a network element that inspects data within the data plane. This inline device performs the traffic shaping, session management and quota management. The advantage of inline policies is that the inspection device can inspect and apply internal local policies many times faster than it takes to trigger a policy to another external element. In addition, service providers can implement more granular policies, as most network elements are limited to applying policies at the layer 3 or the IP classifier.

Upstream versus Downstream Policies

Policy management can also be applied to upstream and downstream traffic. Upstream refers to traffic coming up from the subscriber and going to the network. Downstream traffic originates from the network and flows down to the subscriber. Upstream traffic is, by design, slower than downstream traffic and is considered a limited resource. Both upstream and downstream policies are designed to ensure that no user is starved of access, and are used to prioritize select traffic, such as voice with QoS. They can also protect against certain security attacks, including D/DOS.

The decision to create a policy that is one-sided (meaning applied only to upstream or downstream) is driven by the end goal such as protecting the upstream bandwidth availability or ensuring bandwidth is fairly distributed during peak hours.

Application-centric and Subscriber-centric Policies

A policy infrastructure can also enable service providers to collaborate with their subscribers to allow users to choose the applications that are important to them and to define how they want their Internet service to work. Using application-centric and subscriber-centric policies, the service provider can work to prioritize the applications and incorporate subscriber service level information to meet the requests of their subscribers. Combining the two helps to create rich policies as in the case of a quota-based policy that allows a subscriber unlimited e-mail and Web-browsing, but imposes a monthly quota for volume of bulk file transfers.

Quota/Consumption Policies

Simple quota or consumption caps policies can be employed to limit the amount of data (bits or bytes) that can be downloaded and/or uploaded for a billing period. A subscriber's consumption is measured in both the upstream and the downstream with usage information collected on a per-subscriber and per-application basis.

The per-application information can be used to zero out select applications and/or charge more for other applications. These consumption usage records are then forwarded to a rating and billing system. Consumption-based billing enables service providers to charge for usage over and above a predefined quota, much like mobile phone companies do when users go over their number of minutes in a mobile phone plan.

Consumption billing is often used in conjunction with a quota management system that allows service providers to define actions to take in the event that a subscriber exceeds a quota. These actions could include allowing the subscriber to purchase additional quota, to restrict the allowed bandwidth or to disable their service for the remainder of the billing period.

Quota management is not congestion management—even though it often ends up encouraging users to avoid excessive bandwidth consumption—as peak hour demands will still be the same. The net effect is that people will just use it less during off-hours, which does not address congestion at all. The real benefit of quota management lies in its ability to align the cost of using the network with the subscribers using it, and to capture the network's value in delivering over-the-top services.

This form of traffic management is generally easy to implement and relatively inexpensive to operate as a basic or light duty service. But its use can lead to subscribers misunderstanding how they are being billed, and may even end up alienating the best customers when they are surprised by their higher than expected bills.

Fairshare Policies

Fairshare are “fair usage” policies that include congestion information to enforce a policy in real-time. As the name suggests, Fairshare policies are designed to ensure fairness across users. A policy is applied to heavy users during periods of congestion to allow light users a fair opportunity to use the available bandwidth. This policy usually consists of traffic shaping action and/or traffic prioritization to limit the abusive traffic during the congestion period. By encouraging heavy users to shift their usage to off-peak times, bandwidth can then be divided fairly among users throughout the day. The overall effect is an improved user experience.

Fairshare policies are constructed to be application-agnostic and IP-flow-agnostic. To adhere to the tenets of network neutrality, the base policy is application-agnostic to allow the user freedom of applications and services. Applications may use more than a single IP-flow, so it is equally important that the fairness not be applied at the layer 3 or the IP-flow level but at the aggregate usage of the user.

Fairshare may require service providers to make some infrastructure changes in order to detect and monitor bandwidth use and network congestion. But once implemented Fairshare means

they can better use available bandwidth and eliminate economic inefficiencies in capital and equipment allocation.

Fairshare policies are particularly effective at addressing the bandwidth problems faced by broadband service providers today. Data collected by Sandvine from broadband service providers shows that over a 24 hour period the distribution of heavy users to light users is bimodal, with two percent of the users consuming over 50% of the bandwidth. As more applications adopt natively peer-to-peer data transport architecture, the net effect will be that bandwidth consumption by these applications will continue to grow faster than the subscriber base.

In addition, Sandvine data shows that the demand by the heavy users is inelastic. Based on 15-minute sample periods, 85% of the time the same users consuming the upstream bandwidth in one period would be using it in the next. For the downstream direction, 60% of the time it was the same users.

A Fairshare policy that combines a short-duration quota and network congestion information maintains network neutrality while at the same time helping operators manage their capital expenditure costs.

What about DOCSIS 3.0?

Despite its feature enhancements, the introduction of DOCSIS 3.0 is not the solution to broadband network congestion. DOCSIS 3.0 does incorporate more bandwidth as a means of keeping up with the FTTP rollouts. It also includes some additional mechanisms for traffic optimization whereby traffic can be groomed to channels with the correct transport attributes, such as low latency or high-availability. But even with these updates, bandwidth demand will still continue to outpace bandwidth capacity, as there is only ever a finite amount of spectrum in the cable plant that can be used for data services.

DOCSIS 3.0 may provide a quick fix to the immediate bandwidth crunch; however, the extra capacity will quickly be consumed by even more bandwidth-intensive applications like streaming and transfer of HD quality video—proving once again that network congestion is inevitable. So rather than solving the congestion problem, the introduction of DOCSIS 3.0 only further highlights the need for more traffic optimization.

Optimizing Traffic while Maintaining Network Neutrality

No single traffic optimization technique can fully address all problems relating to network congestion and bandwidth availability. Instead, service providers often lean toward a layered approach, keeping in mind the following:

1. Quota Management is not an effective tool for solving congestion issues because it does not alleviate congestion. Instead, service providers often use quota management as a tool to monetize the over-top-services like streaming HD video, HD YouTube video downloads, etc. In using quota management this way, service providers may seek to change the pricing schemes before these heavyweight services become widely adopted.
2. Fairshare is intended to solve traffic congestion during peak periods. It is a network-centric solution targeted at fairly managing the bandwidth during congestion based upon historical usage patterns of subscribers.
3. Subscriber-centric and application-centric policies should be used to customize the experience for users and make real-time policy decisions and enforcements.

Striking a balance between maximizing the value of the network for all subscribers and maintaining the network neutrality expected by its users is not easy but bandwidth-intensive applications have changed how the game is played for service providers. The minority of users employing these applications impact the overall user experience for the majority of subscribers. When network traffic gets delayed, dropped or degraded, the resulting network congestion has a direct effect on all subscribers. Applying traffic optimization principles levels the playing field for all by alleviating network congestion and ensuring that a limited resource is fairly shared by all users.

Endnotes

¹ http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-243556A1.pdf .

² Network Neutrality: A Broadband Wild West? (Sandvine Incorporated. March 2005)

³ Flow Rate Fairness: Dismantling a Religion, Bob Briscoe. (BT & UCL), ACM Computer Communications Review 37(2) 63-74 (April 2007).

⁴ Charging and rate control for elastic traffic, F. P. Kelly. European Transactions on Telecommunications, 8:33-37, 1997. (Correction by R. Johari & F. Kelly at URL: <http://www.statslab.cam.ac.uk/~frank/elastic.html>).

⁵ Flow Rate Fairness: Dismantling a Religion, Bob Briscoe (BT & UCL), ACM Computer Communications Review 37(2) 63-74 (April 2007).

⁶ Fair Access in Broadband Networks (Sandvine Incorporated. May 2008).