



---

## Contents

Executive Summary	1
Overview	1
History of Traffic Optimization Techniques	2
Application-based Optimization	3
User-based Network Optimization	3
Application and User-based Optimization	4
Reasonable Network Management	4
Evolution of Network Management Models	6
Conclusion	7

## The Evolution of Network Traffic Optimization: Providing Each User Their Fair Share

The challenges facing today's Internet network service providers are a result of technical and business decisions made early in the evolution of public data networks. There is a constant contention between users and operators, applications and networks, and regulation and flexibility.

The application mix on the Internet as a whole is changing as the communications medium evolves from a store-and-forward messaging (email) through client-server bulk (web) to rich media applications. This evolution is driving an increased need for optimization of traffic to yield the best quality of experience for the consumer.

Quota and consumption-based billing are not effective for Internet traffic optimization and congestion management. Broadband service providers who need to manage congestion on their networks must look beyond these to implementing network-enforced user fairness in conjunction with user-selected application optimization. Operators should also be considering the success criteria for fair and reasonable network management that is being contemplated for their networks.

In this paper a brief review of data access optimization techniques is provided, ranging from the earliest days of dialup networking, to a proposed current state of the art involving network-selected subscriber fairness. The success criteria, for fair and reasonable network management is also examined. Additionally, this paper examines user-selected application priority, through to a future-facing dynamic congestion charging model where economic forces are used to align the interests of users, operators, and application writers.

## Executive Summary

This paper provides a brief review of Internet data access optimization techniques, ranging from the earliest days of dial-up networking to state-of-the-art network-selected subscriber fairness and user-selected application prioritization. It examines the success criteria for achieving fair and reasonable traffic management. It also looks ahead to models where economic forces are used to align the interests of users, broadband service providers and application developers.

Sandvine proposes that in the near future, service providers will be implementing network-enforced user fairness in conjunction with a user-selected application optimization system.

Sophisticated network tools that provide detailed visibility into the intent and quality of applications are critical to the network evolution process. This visibility will help ensure quality of service (QoS) for the unique needs of user applications as they compete for access on shared networks.

A key contributor to subscriber satisfaction will be greater transparency for users who are affected by network policies and quotas. This topic is briefly mentioned as it relates to present and future optimization of networks.

Quota and consumption-based billing are not discussed in this document as they are not effective for traffic optimization and congestion management and are best used as alternative charging and service control mechanisms to differentiate tiers of service.

## Overview

Shared networks have been with us since the dawn of the information age. The public switched telephone network (PSTN) has always had more access points (phones) than actual capacity. Operators designed their networks based on realistic peak usage and used mathematics (e.g. Erlang distribution) to help them model for peak usage. As a shared resource, the PSTN was based on call admission control; a call was not admitted to the network unless end-to-end capacity existed to handle it. Today, this is not an efficient use of modern packet-switched network resources as connections last longer and bandwidths are variable.

Data access brought a new complexity to the problem. Access to network resources was no longer determined by slow, human-driven needs such as making a phone call and the number of destinations and paths grew exponentially.

The complexity of managing this intricate and shared network came about with the convergence of other applications onto the packet-switched data network. Each additional application brought unique quality and timeliness demands that added yet another dimension to the network. Traditional admission control was no longer sufficient because the end-to-end network changed dynamically—in the face of mobile IP and other applications competing for the same limited bandwidth.

The 'best effort' service of the past was no longer enough to keep end-users happy. Users expected PSTN quality and guarantees at "best efforts" prices. The complexity of sharing network resources is here to stay, so the challenge is managing it in a fair and efficient fashion. This paper gives a brief history of various techniques and begins to propose a future model that best balances the needs of all parties involved.

## History of Traffic Optimization Techniques

In the earliest days of consumer data access, traffic optimization was given little priority. Congestion was based on the number of ports on a dial-up modem bank and PSTN lines on a phone switch. The earliest dial-up service providers modeled their modem pools with the expectation of a few hours a week of usage per user. Service providers typically supplied access, applications and content using a walled garden approach.

As dial-up access matured, users obtained prioritized PSTN lines for their modems, which caused conflict in dial-up access equipment and in phone switches. Content started to move from proprietary forums and walled gardens to the web. Broadband emerged as a new means to increase capacity and lower cost for access providers.

The emergence of cable (DOCSIS) and phone (DSL) broadband access meant that Internet access became more mainstream and users switched to always-on behavior. The Internet at that time was largely a client-server paradigm where users consumed content that was generated by grass-roots publishing and user-generated content was largely limited to email. This led to the development of asymmetric bandwidth broadband access technologies.

At this same time, technically-savvy users with sufficiently fast computer hardware discovered how to make digital copies of music and learned to 'rip' music from compact discs. Music sharing, previously done by physical copies, grew to meet the digital age and music sharing sites like Napster emerged.

Napster's central server design led to its easy attack and ultimate demise, but the genie was out of the bottle. Distributed music sharing programs such as Gnutella, KaZaa and WinMX became more popular and peak bandwidth per subscriber soared. Bandwidth consumption grew beyond the initial symmetric needs as consumers published as much if not more content than they consumed. Broadband service providers added access capacity as fast as possible to meet subscriber growth and implemented access limits on the TCP port numbers used by these bandwidth-intensive applications.

Consumer data traffic management was born.

The authors of the P2P file-sharing applications responded to the implementation of port-based access limits by moving to dynamic port-based models and later to encryption. At the same time, P2P developers added 'swarming' modes which dramatically increased the ability of a user to download, overcoming the asymmetric nature of the access technologies and the latency of the public Internet. Internet standards anticipated a fixed port number for every type of service, and a new class of intelligent networking equipment products emerged to allow operators to provide policy-based control of the network. Simple policies that involved shaping, session management and priority-based queuing became widely deployed.

As broadband adoption continued to grow worldwide, operators started to leverage their policy-management infrastructure to improve operational efficiencies in areas such as network security. Traffic optimization remained relatively static as long as the operators left sufficient capacity available for consumers to get the content they wished. Then, higher bandwidth video content became available and traffic started to switch back to client-server with the advent of streaming media services such as YouTube®. Not only did this traffic demand a higher bandwidth, it was intolerant of packet loss or temporary reductions in throughput. Higher peak demand for bandwidth meant that a very small number of users could cause quality problems for a wide range of popular applications.

In addition, the rise of mobile data meant increasingly expensive and scarce access resources were being shared by unknown and varying numbers of users. Service providers in this environment started adding user-based management technologies to ensure fairness and provide a consistent quality of experience (QoE) for all users.

## Application-based Optimization

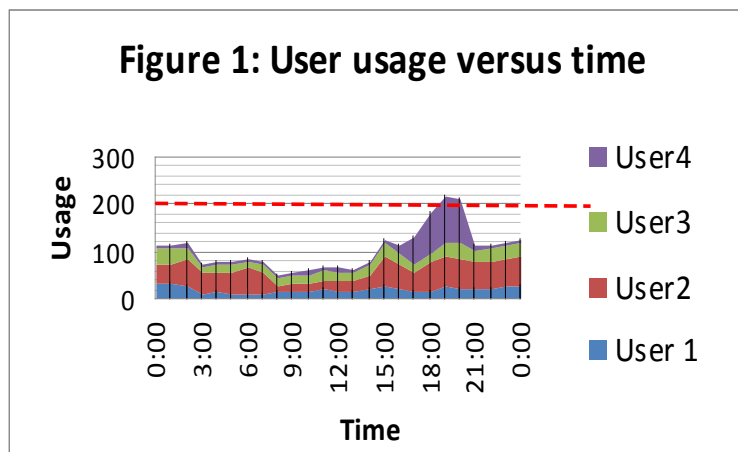
Application-based traffic optimization uses the properties of each network protocol to provide the minimum bandwidth that guarantees acceptable quality. Bulk file transfer applications are given the lowest priority since they are typically non-interactive and long-lived. For example, a one way bulk interactive application such as a file download would be lowest priority, a one-way streaming media like YouTube® may be next in priority and an interactive application such as VoIP would have the highest priority. As the network becomes heavily congested this prioritization becomes important as each application is degraded if it is not prioritized.

Internet standards have anticipated that 'differentiated services' would be offered, where applications 'mark' themselves into the appropriate class based on the priority need of their packets. For example, VoIP marks itself as a high priority given its real-time bandwidth need and a file download marks itself at a lower priority. This provides priority for real-time applications and prevents larger applications from dominating the network. This method, however, is flawed when used in a consumer access application. Broadband access networks (DOCSIS, DSL) do not support 'differentiated services' due to technological limitations. Additionally, differentiated services lead to a fairness issue between subscribers and an incentive to 'cheat', causing the theft of QoS. Application writers sometimes marked their application's packets as the highest priority and this honor system failed.

Service providers have resorted to marking the traffic on behalf of the user, automatically choosing the guarantees that were needed. This application optimization delivers excellent overall quality and subscriber satisfaction.

## User-based Network Optimization

User-based network optimization must be measured over relatively short time periods. Operating on a longer period like daily or monthly does little to affect peak congestion. In figure 1 below, if the available peak capacity is 200 and usage is measured over one hour,



'User4' would be given a lower priority @ 19:00 because she is the top user for that one hour window. If the top user for the 24-hour period was given lower priority (User2), it would not be fair since he was not the heaviest user at the time the congestion occurred, and he would be penalized over the entire 24-hour period and not just during periods of network congestion.

Unlike application-based optimization, user-based optimization provides the operator with a strong tool to give consistent quality and does not lead to network neutrality concerns. However, a strictly user-based optimization model is unfair to the heaviest of users as their traffic is indiscriminately treated regardless of the application they are using (as demanded by net-neutrality doctrine). These users may wish to maintain their overall bandwidth behavior, but control which applications are affected first during periods of congestion.

## Application and User-based Optimization

In this technique, control over access to bandwidth is given to both the service provider and the end-user. The service provider enforces user-to-user fairness allocation and the end-user controls how their individual traffic operates within that allocation. For example, a user may wish to prioritize their VPN access higher than their HTTP, while another user may choose the opposite. During periods of network congestion, the user-based fairness model ensures one end-user's prioritized application does not overly impact another subscriber's.

The simplest method of achieving this goal is to run an application-based optimization instance for each user, providing weighted fair queuing within their traffic envelope. The service provider would then run a separate set of queues which would balance the traffic between users.

To increase subscriber satisfaction through personalization of service, the service provider may wish to give each user more control over their own priorities. This may involve a 'quota' of QoS points or a web page which gives specific weightings per application or per application class. There would be no change in billing plans to operate this service, which makes it very feasible with today's technology and consumer education level.

This scheme is clearly the best because it provides a network-neutral and consumer-transparent sharing of network bandwidth resources.

## Reasonable Network Management

For any traffic optimization technique to be successful and provide fair and reasonable network management, it needs to ensure it meets 5 key success criteria:

1. Narrowly Tailored
2. Proportional and have reasonable affect
3. Legitimate and have a demonstrable technical need
4. Transparently disclosed
5. Auditable

### Narrowly-Tailored

All networks have variation in usage patterns, whether by time of day, by geography, by user demographics, or by other factors. As a consequence, oversubscription and quality of

experience are non-uniform across the network. A properly constructed network management plan takes this into account, and focuses as narrowly as possible on the problem to be solved. It does not try to force a one-size-fits-all solution into all areas at all times.

In any shared access network environment, there are several areas of narrowly-tailored that might be considered:

- 1) Peak network demand and capacity constraints in both the access and backhaul networks
- 2) Subscriber density per constrained resource
- 3) Subscriber demographics per constrained resource
- 4) Unforeseeable events

A reasonable network management practice takes these factors, and more, into consideration. It applies itself differently, or not at all, depending on the conditions which are currently present. For example, a network management practice might be self-tuning, and disable itself when no congestion is present. It might operate differently when congestion is present on a single user, versus a single constrained access resource, versus a single constrained backhaul resource.

A successful network management practice will narrowly tailor itself to the problem at hand at the time it is needed. It will not apply in a broad fashion across the broad average of a network.

### **Proportional and Reasonable Effect**

The impact of a reasonable network management policy must be proportionate to the problem being solved. It would be considered unreasonable by most to take a subscriber causing 15% of the congestion on a network, and manage their bandwidth to 1% of peak rate. The network management policy needs to ensure that the policy applied to subscribers is proportional to the level of impact they are having on the constrained resource. Network management plans need to take into account the concept of proportional affect and response.

### **Legitimate and Demonstrable Technical Need**

The operator must have a legitimate and demonstrable technical need for the network management practice. The architecture strengths and weaknesses of the various access technologies provide the majority of the technical needs for network management. Additional technical needs arise due to network architecture outside of access network specifications, for example implementation-specific details of various access infrastructure vendors, backhaul and core network architecture.

To be successful, a network management practice must be described in such a way that the technical need that is being addressed is clear, and that the practice that is employed is designed to address this need and nothing more.

### **Transparent Disclosure**

The operator must make the material information required to understand the network management policy publicly available to those impacted by it. The disclosure should be sufficient for a consumer to develop an informed opinion on whether the practice will affect

them, which applications might be affected, when they might be affected, and what the impact might be, including impact to speed, latency and overall quality of experience.

Disclosure might take many concurrent forms. The most popular include network management FAQ web pages, notices included in billing material, acceptable use policies, terms of service, etc.

### **Auditable**

Owing to the public scrutiny of capital investment in networks, and network management policies, it becomes important for operators to be able to demonstrate that the above criteria were indeed met.

On audit, an operator should be able to provide:

- 1) What the technical need was that caused the creation of the network management policy
- 2) What affect the policy had on subscriber experience
- 3) How the policy was disclosed to the end-user
- 4) How the policy was narrowly tailored to take into account network and time variances, for example.

In addition, the audit should be able to demonstrate the above criteria were met using technical results. These results might include information on the subscriber experience for the typical subscriber in typical network locations.

## **Evolution of Network Management Models**

Sandvine believes that once traffic optimization reaches the stage where both the needs of the end-user and the needs of the operator are effectively balanced, the traffic optimization model will evolve once again.

All of the traffic optimization techniques mentioned in this paper have assumed that an appropriate level of investment is made in the network, without subscribers really having any input on what that level may be. An assumption exists that oversubscription is managed to 'acceptable' levels, and these traffic optimization techniques are used to provide the highest, most consistent quality for a given oversubscription. The issue of managing oversubscription to 'acceptable' levels is amplified as internet connectivity shifts to wideband access and the peak to trough ratio increases. A user is not able to 'pay' an additional amount for a higher guarantee, and in turn a network provider is not incented to provide higher guarantees that are required by competitive forces.

In a dynamic congestion charging scheme, each packet has an effective real-time auction for how much it will 'pay' for the available capacity. In practice, this is done by giving each user a number of 'points' for packet priority in each interval, and letting them spread them however they wish. Premium service plans get more 'points' that may be used.

Subscriber fairness is still enforced as in the models described above, but now we have an alignment of interests: the access provider is incented to provide all the QoS - sensitive

bandwidth their users will pay for, users are not incented to 'cheat' one another since it 'costs' them, and application writers are not incented to maximize bandwidth.

For simplicity of messaging and communication, an access provider will probably launch this service with 'cost' assigned by time of day. This would be similar to the weekend/evening minutes plans the mobile carriers use, and consumers will have an easier time understanding the 'cost'. The user would be able to mark each application / packet according to priority ranging from 'isochronous high guarantee' to 'scavenger class', and pay nothing for the scavenger traffic. This in turn makes the traffic optimization and capacity planning of the network service provider much more direct.



Efforts are underway in the IETF, one of the standards bodies which govern the Internet, to create a model for charging for congestion based on feedback of congestion caused. This is called 'congestion exposure' (ConEx).

## Conclusions

Internet traffic optimization has come a long way from the early days of dial-up access, both in terms of the requirement for it, and in terms of the complexity and technical effectiveness of the methods used.

Today's state of the art network management plan provides network-enforced inter-user fairness, and allows end-users the ability to prioritize their applications as they see fit. Access network operators should start messaging this duality of methods and the benefits to their subscribers.

Without regard to the specific network management practice employed, policies must be narrowly tailored, must be proportional and reasonable, must be designed to address a legitimate technical need, must be transparently disclosed, and must be auditable.

Network policy control is required to create a network management practice that spans multiple devices, and multiple access technologies.

Strong reporting and business intelligence is required to be coupled to the network management practice to understand demand, capacity, and user experience.

The next logical step in the evolution of consumer data services is going to be a fundamental shift in the charging methods used. This will invoke economic means to incent operators to build the best network, and to incent users to use it in the most efficient fashion. In order that operators and end-users reap the benefits of this new model, supply will need to always slightly exceed demand, and free market forces will prevail. Any business model which is founded solely in arbitrage of costs will disappear, leaving a stable, long-term equilibrium.