



---

## Contents

Executive Summary.....	1
Introduction.....	2
Congestion Management Principles.....	2
The Nature of Congestion.....	2
Quota Management Levers.....	5
Sliding Window Usage Measures.....	6
Reporting, Business & Network Intelligence.....	8
Conclusion.....	9

# Quota Management Does Not Solve Congestion - Traffic Management Does

## Executive Summary

Ever growing demand for broadband bandwidth in both fixed and mobile networks is leading many operators to introduce data caps in place of unlimited traffic service plans, or as alternative plans. They are aiming for a market segmentation that better aligns revenue with network cost.

The introduction of such data caps (herein referred to as quota management) is often motivated not only by the desire to maximize revenue via better market segmentation, but also to dampen traffic demand in networks that, at certain times or in certain locations, experience over-subscribed demand and become congested.

While quota management will dampen some demand for bandwidth, it does not make for effective congestion management; nor does it help the operator significantly in judging the business case for extra network investment.

This paper explains the weaknesses of quota management on both those counts and describes the kind of traffic management tools that do address congestion cases effectively, as well as provide the intelligence needed for investment decisions.

The clear conclusion is that operators are best served by deploying quota management capabilities primarily for revenue-maximizing service plan segmentation while using congestion-targeting traffic optimization products to manage congestion events and support network investment decisions.

Sandvine provides the leading congestion management solution with its Fairshare Traffic Management product and enables quota-based service plans via its Usage Management product suite.

## Introduction

Ever growing demand for broadband bandwidth in both fixed and mobile networks is leading many operators to introduce data caps in place of unlimited traffic service plans - or at least as alternative plans aimed at a market segmentation that better aligns revenue with network cost.

The introduction of such data caps (herein referred to as quota management) is often motivated not only by the desire to maximize revenue via better market segmentation but also to dampen traffic demand in networks that at certain times or in certain locations experience over-subscribed demand and become congested.

It will clearly be the case that a user who is moved from an unlimited traffic plan to a fairly restrictive capped traffic plan will be more cautious in their use of the network and hence reduce their traffic demand. However much of this reduction could be during times of ample spare capacity in their part of the network. This neither benefits the experience of other users nor does it reduce the need for network investment. At the same time, the reduction in demand at times, or on nodes, where congestion occurs may be small in relation to the need for intervention. Furthermore, there will be many cases of users at the start of their billing cycle who are unconcerned about breaching their quota at that time and hence do not curb their traffic demand at all.

It is important therefore to understand the typical peak period or congestion scenario and how the levers that are available to operators via network policy control systems can best be used to manage congestion. This should be done, context permitting, against the background of five principles for congestion management which Sandvine recommends.

## Congestion Management Principles

- Narrowly-tailored
  - Actively manage only where congestion exists and when congestion is causing Quality of Experience (QoE) issues for a large number of subscribers
- Proportional and reasonable effect
  - Policy should have an effect on subscribers or applications that is proportional to the effect the user or application is having on the network
  - Policy applies the smallest reasonable intervention to alleviate congestion and improve the QoE for the majority of subscribers
- Legitimate and demonstrable technical need
  - Management is effective in achieving its targeted goals and the maintenance of subscriber QoE
- Transparent disclosure
  - The policies applied allow the operator to disclose its traffic management policies in a simple and understandable manner
- Auditable
  - The management solution allows the service provider to demonstrate that the above requirements were met through its auditing and reporting capabilities

## The Nature of Congestion

There are three main forms of traffic management appropriate to congestion conditions:

- Fairness of access to bandwidth, between like-subscribers (as distinct from equal access per IP flow, for example)
- Constraint or lower priority for applications that are not time-sensitive (such as P2P)
- Constraint or lower priority for “heavy users” who are taking, or who have taken, more than their “fair share” of network resources.

Most subscriber-aware access network systems automatically provide fairness between subscribers but without special treatment of either applications or heavy users. Management of this kind alone will not prevent high simultaneous use of P2P and time-sensitive applications from resulting in poor quality of experience; neither will it prevent heavy users’ behavior from adversely impacting other users’ quality of experience. So it is worth looking at the nature of the two phenomena of P2P and similar applications on the one hand and heavy users on the other.

Sandvine's most recent Broadband Phenomena reports<sup>1</sup> show an increasing proportion of real-time, time-sensitive applications, especially in peak periods. Nevertheless P2P remains a significant contributor to peak demand, in particular in certain regions and in the upstream. Figure 1 below, illustrates this with peak period traffic for the CALA region from the Mobile Internet report.

## Caribbean and Latin America-Network Aggregate Peak Hours

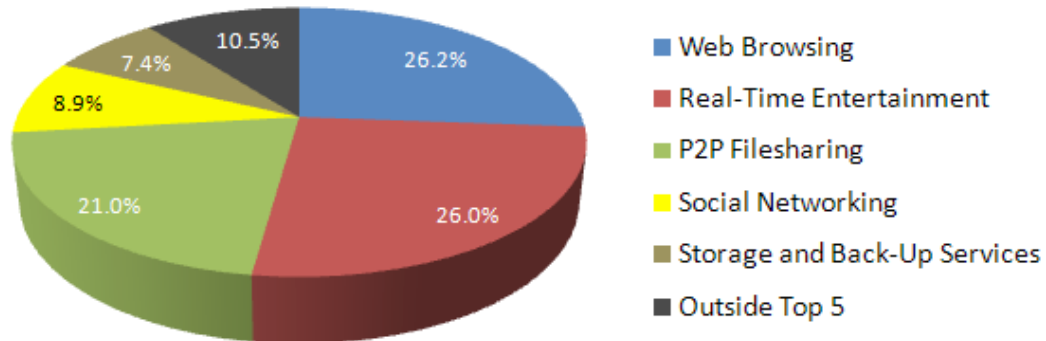


Figure 1 Share of Peak Period Traffic, CALA Region, Sandvine Mobile Internet Report, March 2010

Figure 2, from the 2009 report on fixed network usage, shows significant use of P2P in particular in the upstream.

## 2009 Average

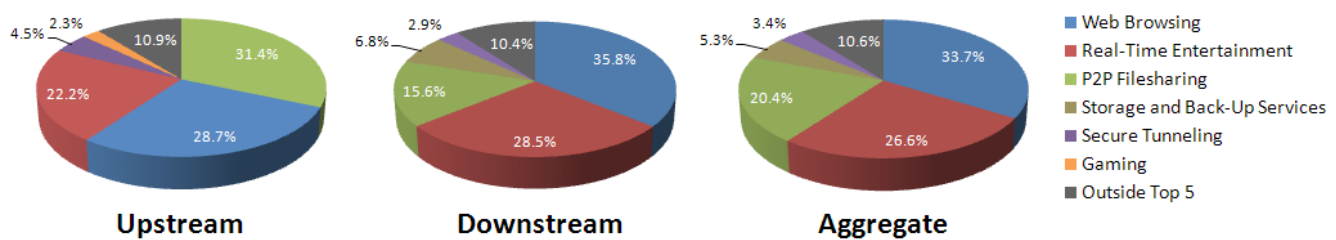


Figure 2 Share of total fixed network traffic, Sandvine Global Broadband Phenomena Report, Oct 2009

Figures 3 and 4 show how P2P continues to be the dominant traffic type among top users of both mobile and fixed networks. As we shall argue below, the top users by overall traffic (e.g. per month) are not necessarily the ones who warrant some intervention in managing congestion. However these results at least hint that high traffic users generally may be high P2P users and hence one may be able materially to impact peak period demand by targeting only the P2P (or non time-sensitive) traffic of only the users contributing most to congestion.

## Europe - Top Subscriber Aggregate Traffic (Peak Hours)

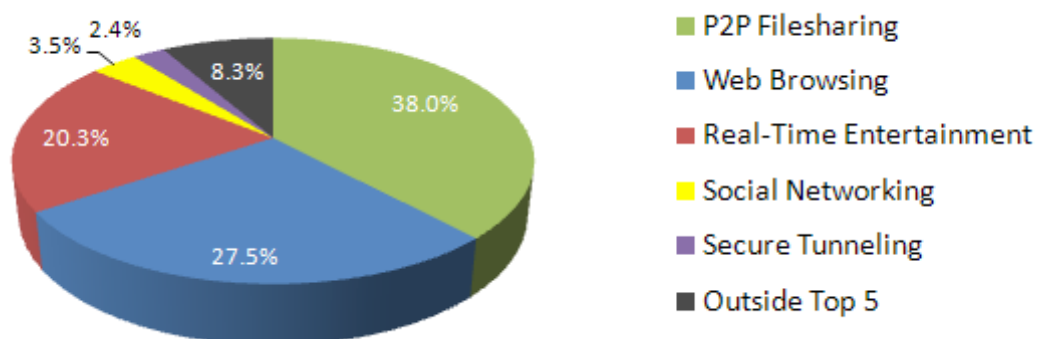


Figure 3: Share of Top Subs Peak Period Traffic, Europe, Sandvine Mobile Internet Report, March 2010

1 Global Mobile Internet Phenomena Report, March 2010, and Global Broadband Phenomena Report, October 2009.

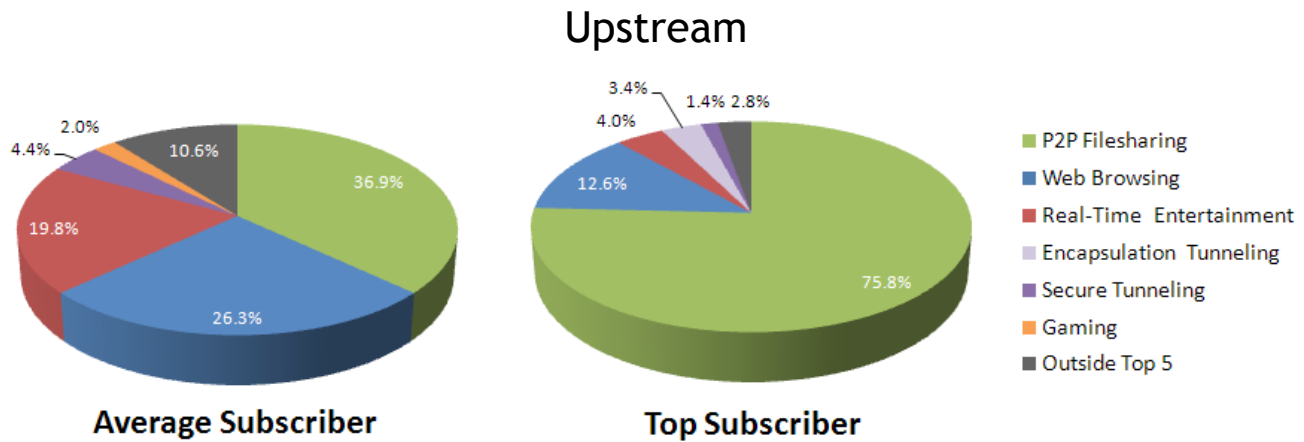


Figure 4 Top Subs' share of upstream fixed network traffic, Sandvine Global Broadband Phenomena Report, Oct 2009

At a high level, the importance of heavy users in broadband networks is indicated by the following observations from the Sandvine 2010 Mobile Internet report and other, sampled customer networks:

- For Mobile upstream the top 10% of users account for 75% of the traffic
- The top 1% of mobile subscribers in Europe account for more than 30% of aggregate mobile network
- In a typical fixed network in the peak period, the top 2% of users in that period account for about 50% of the traffic in the downstream
  - The upstream behaviour is even more skewed towards the heavy users
  - Such, short-term heavy users, suitably targeted, offer the best potential for congestion management that accords with the principles described above.

Figure 5 shows the cumulative usage percentile by subscriber percentile for networks in Sandvine's 2010 Mobile Broadband Phenomena report.

## Cumulative Usage Percentile by Subscriber Percentile (Monthly Aggregate)

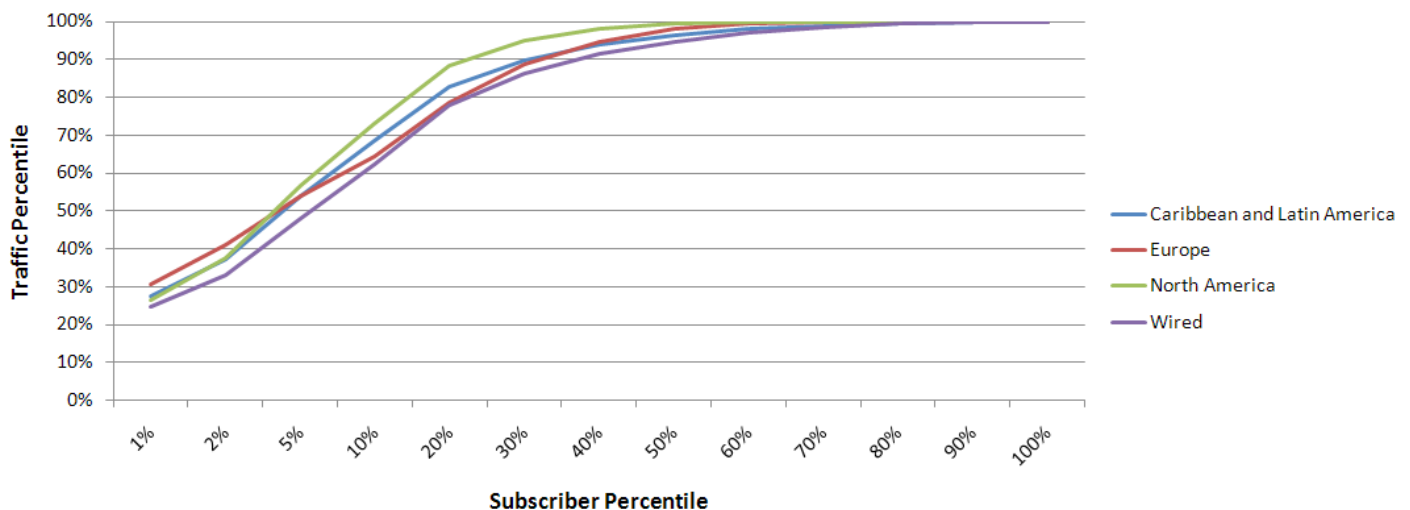


Figure 5: Monthly Usage Percentile by Subscriber Percentile: Mobile Broadband

## Quota Management Levers

Quota management in the form of monthly data caps generally results in highly restricted bandwidth once the cap is exceeded, often so restricted that the connection is no longer useful for normal purposes. There may be an option to top up for an additional data allowance or, instead of any restriction, there can be simply a metered charge for traffic above the initial cap.

As noted already, while one motivation for quota management is to curb traffic volume growth, the main one is to maximize revenue and to better match revenues to costs. In practice a monthly data cap with revenue implications will need to be one which the user generally does not exceed during the month. Otherwise the user is probably in the wrong service plan and will be frustrated by the messaging and costs of their frequent crossing of their traffic limit. Consequently, a sensible quota management scheme will see only a small minority of users being in excess of their traffic limit. Furthermore, any policy constraining such users is not necessarily affecting those users contributing the most to congestion in the network. It simply singles out those who have used more traffic than their service plan allowed.

This is illustrated with a simple example in Figure 6 below.

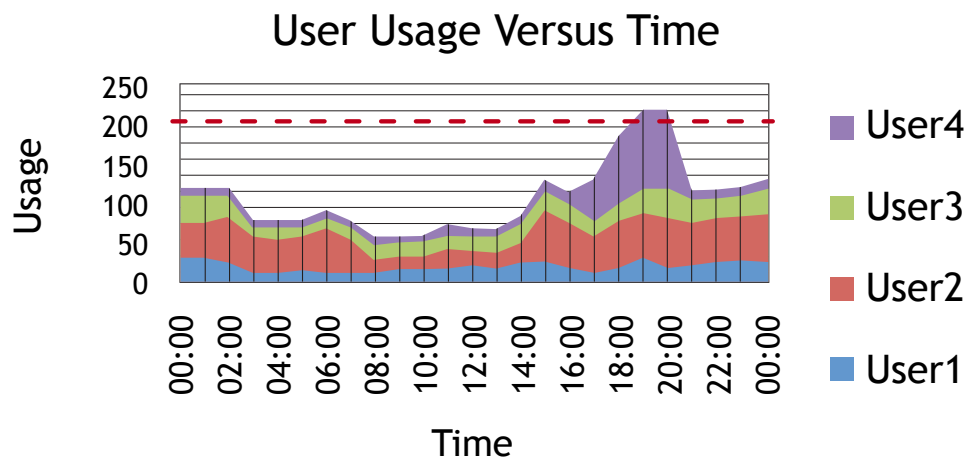


Figure 6: Long term usage having no bearing on congestion management

In this example, if the available peak capacity is 200, it is most likely that we want 'User4' to be given a lower priority at 19:00. Yet User 4 has consumed the least amount of traffic in the last 18 hours. Nevertheless they are contributing most to congestion, significantly more than User 2 who has the largest long-term usage. The most effective predictor of congestion contribution is not long term but very short term usage data.

Another issue with long term usage quotas is that there will also be many users who, comfortable in the knowledge that they are not likely to exceed their monthly allowance, do not curb their demand at all during periods of congestion in the network. This is especially so at the start of the monthly allowance cycle, when the prospect of exceeding the allowance seems slight.

Some data caps are based on the calendar month. This has the result that at the start of the month hardly anyone feels inhibited by their allowance and hence there is no effective constraint at all. Where the data cap applies to the billing period, and where those period starting dates are evenly distributed across the month, then such demand dampening should also be more evenly distributed across the month. Nevertheless, it will tend to be only those coming towards the end of their billing cycle who pay much attention to their data cap, while the majority behave much as if they had no cap.

In sum, any restriction placed on users when they exceed their quota is there to match overall usage to the user's service plan. The only direct intended consequence is to induce the user to buy a top up or upgrade their plan. The constraint is unrelated to any congestion in the network and should be a relatively rare event. By the same token, the large majority of users under capped service plans who are well within their traffic limits may behave much the same as they would without the cap and in particular may include short-term heavy users who contribute disproportionately to congestion. Quota Management, which is appropriately calibrated for segmenting the market into popular service plans, will thus not be effective for congestion management.

## Sliding Window Usage Measures

We have noted the potential for managing “heavy user” traffic as an effective means of reducing peak traffic load or of improving the experience of normal users, either in conjunction with general (aggregate) management of P2P and similar applications, or on its own. At the same time, we have just noted that a user who is at or near their quota limit is not a good definition of a heavy user for congestion management purposes. So what usage measure is appropriate for congestion management?

We need to avoid the test for heavy user status being calendar-dependent (e.g. a user flips from being a heavy user at the end of one monthly period to being an extremely light user as soon as the calendar-based counting cycle starts again). To do this, it is necessary to use sliding (or “trailing”) windows for measuring usage. That means always counting the period back from the time that any policy might be applied. As one moves forward by one period in time, the sliding window count adds the traffic from the latest historic period and subtracts the traffic from the oldest period included in the previous count.

For example, suppose we use a 5 period sliding window and on some calendar scale we are just entering period #12346, with the sliding window traffic count = T. T is thus the sum of traffic in periods #12341-12345. As we enter period #12347, T is adjusted by adding the traffic for period #12346 and subtracting the traffic for period #12341.

Two important things to note about a sliding window measure:

- a) The measure goes down as well as up. Unlike a usage quota it does not only increase; and it may well never go above a specified threshold.
- b) There is no calendar-dependent, arbitrary switch in values (from high to zero). A persistently high user will have a continuously high sliding window usage.

Given that we need to use a sliding window measure (which is separate from any usage management quota), what periods should be used for the purpose of heavy user congestion management policy?

Table 1, below, is based on peak period downstream data for a fixed network. Users in the network have been ranked according their usage in that hour and then again by a sliding window of 1 day, 1 week, and 1 month back from that time. The percentage figures in the table are the correlation coefficients for those rankings against the 1 hour ranking.

	RxHourRank
RxHourRank	100.00%
RxDayRank	44.03%
RxWeekRank	39.20%
RxMonthRank	38.79%

Table 1: Fixed Downstream Sample: Ranking by traffic - correlation of sliding window periods vs. single peak hour

This table tells us that the daily, weekly, and monthly ranking give broadly similar results in predicting 40% of the top users in the peak hour and missing 60% of them.

Work on another sample set (also for fixed network downstream traffic) examined the traffic of the top 2% of users in periods ranging from 15 mins to 24 hours. The results of this analysis are shown in Table 2, below. They indicate that the shorter the historical period, the more accurate the prediction of high usage.

Indicator	Target	% of target periods found with Indicator
Top 2% users ranked by usage on the previous day (ending midnight)	High use 15 minute periods where the high use threshold catches 2 % of users over the 4 peak hours	25%
Top 2% users ranked by usage over the trailing 24 hours		48%
Trailing 60 minute high use where high use threshold catches 2 % of users in peak hours		46%
Trailing 30 minute high use where high use threshold catches 2 % of users in peak hours		63%
Trailing 15 minute high use where high use threshold catches 2 % of users in peak hours		79%

Table 2: Declining efficiency of longer term usage triggers in the downstream

A sample set with some visibility of behavior by protocols (also for the downstream) for time ranges between 1 hour and 5 minutes shows very high rates of correlation (approx. 80%) between the top users in the last period and those in the next period for Gnutella and NNTP protocols. In those cases the correlation is not sensitive to the time period chosen (within this limited range of times). However, this changes radically if one wants to be protocol blind. In that case, the correlation drops to 20% for the one hour period, compared with approx. 60% for 10 and 15 minute periods.

The top 2% users by volume, in peak time 15 min periods, account for 50% of all peak time traffic and 40% during those periods when they are above the threshold bandwidth. If one can predict around 75% of such cases, this suggests that up to around 30% of downstream peak period bandwidth use could be successfully targeted via a well chosen heavy user policy. Policy modeling and practical experience indicate that peak downstream demand could be reduced by 15% or more in fixed networks by such policies. The effect of upstream policies can be significantly greater. In the upstream heavy use tends to be more persistent and hence the longer term sliding window usage measures do better at predicting heavy users than they do for downstream.

In the downstream, in Table 2, we have seen prediction of 15 min period high use increase from 25% accuracy based on the last full day's traffic, through to 79% based on the last 15 min data. That suggests that the shorter period used, the better - a far cry from using usage management monthly quotas.

There is some evidence that the accuracy does not improve significantly for periods below 10 minutes. In any case, to use only a very short period as a condition for some penalizing policy runs the danger of effectively penalizing users for simply using the bandwidth that has been promised in their service plan. The accuracy of short term usage indicators has to be set off against a fairness criterion that the usage needs to be in some sense extraordinary or excessive in order to warrant special management.

Many operators will feel that a 15 min usage trigger for policy alone can indeed be fair, if the trigger value is set high enough - indicating sustained very high usage during that period - and especially if the user's experience will only be impacted if there is congestion in the network.

Other operators will feel that only persistently heavy users should be subject to special measures and may therefore want a 7 day sliding window measure (for example) as a policy condition. To accommodate this, one could combine a top 2% of peak period 15 min usage trigger (to predict impact in the next 15 min period) with a top 5% or even 10% on a 7 day sliding window (for a persistently heavy user condition). This will obviously reduce the number of high use cases impacted but not as much as simply using the top 2% of the 7 day window. The tradeoff between loss of accuracy and greater perceived fairness is for the operator to judge in the light of the amount of bandwidth that needs to be constrained or lowered in priority in order to meet normal users' performance expectations.

Maintaining a short-term policy condition in all cases has the advantages that:

- Subscribers who breach some longer term “fairness” threshold but are not exhibiting short-term high usage behavior will - through the short-term usage condition - be exempted from policies which could have otherwise worsened their current experience;
- Policy actions, especially when also conditional on known congestion status, can be limited to cases where they can be expected to make a significant difference and hence the reporting of those actions contributes greatly to both business intelligence (how are my users being impacted?) and network intelligence (where and when are the events that cause congestion management actions?) .

## Reporting, Business & Network Intelligence

Operators concerned about the effect of congestion on their users need not only to manage traffic to ensure fairness and to improve the user experience; they also need to know where the congestion is occurring, how much they are policing both individual users and classes of user, and the impact of traffic management actions on the congested parts of the network.

If quota management alone were relied upon for congestion management, the operator would be limited to looking at usage statistics and the incidence (either per class or by sub) of over-quota penalty actions. None of this would be mapped to congestion events or congested elements in the network. At the network level, the operator is only able to track measures such as total utilization per resource and hence to base network investment decisions on indicators of persistent under provisioning of capacity based on such measures.

Using congestion management tools separate from quota management, targeted at congestion events, improves the normal user’s experience. It also allows the operator to segment traffic usage between high and low priority classes and then to focus on the demand for high priority traffic and related measures of the normal user’s quality of experience as the appropriate indicators for network upgrades.

With congestion dependent policies potentially impacting both traffic throughput and its priority classification, it is important to provide visibility of the relationships between congestion management policies and:

- a) On the network side - per resource:
  1. Total traffic
  2. Users and traffic per traffic class
  3. Quality measures per user per class
- b) On the subscriber side:
  1. Policy state and history per user
  2. Policy history per class

Additional value can be provided through enhanced data analysis that can superimpose network upgrade events or policy change events on the history; or that can project trends into the future to forecast network upgrade dates or estimate the impact of alternative policies.

The right network intelligence tools allow one to move away from crudely provisioned capacity which will result in variable and often poor user experience to capacity much more intelligently targeted at providing the appropriate quality of experience per user tier, per application, and per class of usage (normal vs. beyond normal behavior). Network investment decisions should not be based on a single measurement, such as link utilization, but should be multi-dimensional, taking into account measurements including:

- Link utilization
- Quality of experience per tier of subscriber
- Number of subscribers being managed
- Quality of experience of subscribers who are managed and not managed

## Conclusion

In both fixed and mobile networks, quota management in the form of data caps (or traffic quotas) is becoming popular for either optional or compulsory service plans in order to help align revenue with network cost. Operators introducing these may be at least partially motivated by the desire to dampen general traffic demand and hence contain network cost. This paper has shown that quota management techniques are not effective in the management of traffic and users in real-time congestion events. They also contribute nothing to the kind of network intelligence needed for justifying network investment in order to positively impact subscriber quality of experience, resulting in the improvement of customer loyalty and increased ARPU.

Network policy control products aimed at congestion management need to:

- a) Be application-aware: to either exempt or include applications, as appropriate;
- b) Use sliding window traffic usage measures that can fall, as well as rise, and bear no relation to service plan or calendar-related traffic quotas;
- c) Map policy actions wherever possible to congested network resources and to congestion conditions;
- d) Have powerful reporting and analysis tools to provide the necessary feedback loop on congestion management in both the customer relationship and network domains.

Sandvine's Fairshare Traffic Management product has been designed specifically to address the issues raised in this paper. It equips the service provider with the most accurate means of targeting (in real-time) the users most contributing to congestion, changing their priority or capping their bandwidth only when congestion is present, and providing the visibility on both network usage and policy events which is key to network planning as well as to the audit and transparency of traffic management policy.

Sandvine's Usage Management product suite, in turn, addresses the revenue challenge of all you can eat service packages in the face of the burgeoning traffic growth on the one hand and service segmentation purely by maximum bandwidth on the other. Competitive pressure and user demand imply the need for higher bandwidth offerings, but this can induce many users to drop their tier and hence spend less.

Sandvine Usage Management products can turn that challenge into a revenue opportunity by introducing metered access options, which effectively segment high volume usage subscribers from low volume ones -irrespective of their maximum bandwidth. These options can be based either on volume or time, and either on gross usage or segmented by content - with zero-rated options in either case.