

The Effect of ISP Traffic Shaping on User-Perceived Performance in Broadband Shared Access Networks

Kyeong Soo Kim^{a,*}

^a*College of Engineering, Swansea University, Swansea, SA2 8PP, Wales United Kingdom*

Abstract

Recent studies on the practice of shaping subscribers' traffic by Internet service providers (ISPs) give a new insight into the actual performance of broadband access networks at a packet level. Unlike metro and backbone networks, however, access networks directly interface with end-users, so it is important to base the study and design of access networks on the behaviors of and the actual performance perceived by end-users. In this paper we study the effect of ISP traffic shaping using traffic models based on user behaviors and application/session-layer metrics providing quantifiable measures of user-perceived performance for HTTP, FTP, and streaming video traffic. To compare the user-perceived performance of shaped traffic flows with those of unshaped ones in an integrated way, we use a multivariate non-inferiority testing procedure. We first investigate the effect of the token generation rate and the token bucket size of a token bucket filter (TBF) on user-perceived performance at a subscriber level with a single subscriber. Then we investigate their effect at an access level where shaped traffic flows from multiple subscribers interact with one another in a common shared access network. The simulation results show that for a given token generation rate, a larger token bucket — i.e., up to 100 MB and 1 GB for access line rates of 100 Mbit/s and 1 Gbit/s, respectively — provides better user-perceived performance at both subscriber and access levels. It is also shown that the loose burst control resulting from the large token bucket — again up to 100 MB for access line rate of 100 Mbit/s — does not negatively affect user-perceived performance with multiple subscribers even in the presence of non-conformant subscribers; with a much larger token bucket (e.g., size of 10 GB), however, the negative

*Tel.: +44 (0)1792 602024.

Email address: `k.s.kim@swansea.ac.uk` (Kyeong Soo Kim)

effect of non-conformant subscribers on the user-perceived performance of conformant subscribers becomes clearly visible because the impact of token bucket size and that of token generation rate are virtually indistinguishable in this case.

Keywords: Traffic shaping, access, Internet service provider (ISP), user behavior, user-perceived performance, quality of experience (QoE)

1. Introduction

The practice of shaping subscribers' traffic by Internet service providers (ISPs) has been under intensive study; for example, the effect of ISP traffic shaping on various packet-level performance with its detection and policies has been investigated based on actual measurements and mathematical/simulation analyses in [1], [2], [3], and [4], which provides a new insight into the actual performance of broadband access networks at a packet level.

Unlike metro and backbone networks, however, access networks directly interface with end-users, so it is important to base the study and design of access networks on the behaviors of and the actual performance perceived by end-users [5]. The major goal of our study in this paper, therefore, is to investigate the effect of ISP traffic shaping on user-perceived performance, i.e., the quality of experience (QoE), and thereby to provide ISPs further insights into the design, deployment, and operation of the next-generation access networks from end-users' perspective. Because in access networks the average rate of a service is determined by a service contract between a subscriber and an ISP (e.g., through subscription tiers), which is then controlled accordingly by the token generation rate of a token bucket filter (TBF) used in traffic shaping, and the peak rate is usually determined by the underlying access technology (e.g., line rates of digital subscriber line (DSL) and cable Internet), we put our major focus on the effect of the token bucket size of a TBF on user-perceived performance given the token generation rate and the peak rate of an access link.¹

The current study was specifically triggered by the results from recent

¹The token generation rate and the token bucket size correspond to the maximum sustained traffic rate (MSTR) and the maximum traffic burst in the data over cable service interface specifications (DOCSIS) media access control (MAC) and upper layer protocols interface specification [6], respectively.

investigations of “PowerBoost” in [1] and [2], the feature present in some cable broadband networks that enables sharing of unused capacity by giving customers extra bursts of speed whose duration is controlled by the token bucket size. In this paper we extend those investigations of the effect of the token bucket size on packet-level performance at network/transport layers to those on user-perceived performance at application/session layers based on the research framework which we proposed for the clean-slate design of next-generation access networks [7, 5].

Through the investigation we answer the following key questions:

- **Subscriber level:** If we consider a single subscriber under traffic shaping in isolation, what is the minimum token bucket size providing user-perceived performance *nearly equivalent*² to those of a subscriber under no traffic shaping for a given token generation rate and a mix of traffic flows?
- **Access level:** In a shared access, where shaped traffic flows from multiple subscribers interact with one another, what is the effect of the token bucket size on user-perceived performance of all subscribers? Specifically, how many subscribers can be served with user-perceived performance nearly equivalent to those of a subscriber under no traffic shaping in a dedicated access for a given token generation rate and a mix of traffic flows as well as a token bucket size suggested by the subscriber-level investigation?

It is this *trade-off in traffic shaping* between the performance at subscriber and access levels that interests both subscribers and ISPs. Considering the bursty nature of traffic flows at multiple layers, e.g., user behaviors and variable bit rate (VBR) encoding at the application/session layers and transmission control protocol (TCP) flow and congestion controls at the transport layer, one can expect that at the subscriber level, the user-perceived performance of a subscriber under traffic shaping with loose burst control would approach those of a subscriber under no traffic shaping when the token generation rate is equal to or greater than the long-term average rate of combined traffic flows. At the access level, on the other hand, it is likely that the loose

²The term “nearly equivalent” is formally defined based on multivariate non-inferiority testing in Sec. 3.3.

burst control at the subscriber level negatively affects the performance of other subscribers, especially when there are non-conformant or mis-behaving subscribers.³

Note that the traffic shaping and related issues (e.g., multiplexing and scheduling of shaped flows) have been extensively studied in the context of per-flow/connection traffic shaping and based on packet-level measures since the introduction of the “leaky bucket” method in [8].⁴ In [9, 10, 11, 12], the TBF and their analyses with various statistical traffic models are studied. In [13, 14], the dimensioning of TBF parameters through the notion of a linear-bounded arrival process (LBAP) is investigated for aggregated voice over Internet protocol (VoIP) and long-range dependent (LRD) traffic. In [15], the performance trade-off of traffic shaping between access control queueing and network queueing is studied based on the spectral analysis technique, while in [16], the characterization of LRD traffic regulated by leaky-bucket policers and shapers is studied using the modified Allan variance (MAVAR) for the LRD estimation and spectral analysis of the regulated traffic. As for scheduling of shaped traffic flows, the end-to-end delay bounds and the buffer space requirements of various scheduling disciplines are well summarized in [17].

The results from these studies suggest that given a token generation rate, allowing large bursts through a large token bucket size improves the packet-level performance of an individual flow, while multiplexing of those shaped flows would increase the deterministic bound of end-to-end packet delay. These works, however, are not done in the context of ISP traffic shaping, where multiple traffic flows with different service types are shaped together by a single TBF, and do not take into account user behaviors in traffic generation and performance perceived by end-users. To the best of our knowledge, our work in this paper is the first attempt to systematically assess the effect of ISP traffic shaping on user-perceived performance with user-behavior-based traffic models at both subscriber and access levels.

The rest of the paper is organized as follows: Section 2 provides an

³Subscribers who consistently generate traffic whose long-term average is higher than that of the service contract (i.e., the token generation rate) are called *non-conformant* or *mis-behaving* subscribers in this paper.

⁴The leaky bucket algorithm described in [8] is basically the same as the token bucket algorithm. We use the terms “leaky bucket” and “token bucket” interchangeably in this paper.

overview of the current practice of ISP traffic shaping and its major issues. Section 3 describes the methodology we adopt for this investigation with details of experimental setup and a comparative analysis framework. Section 4 presents the results of experimental investigation of the effect of ISP traffic shaping at both subscriber and access levels. Section 5 concludes our work in this paper.

2. Overview of ISP Traffic Shaping

2.1. Current Practice

Traffic shaping was originally devised for connection-oriented networks to regulate an *individual flow* per traffic conformance definition negotiated during a connection admission control (CAC) at a user-network interface (UNI) [8, 18]. It is now used by ISPs to regulate *combined flows* from a subscriber in a different context of connectionless IP networks: Because there are no CAC procedures used at the UNI in the current IP-based networks, ISPs base their traffic shaping on service contracts with subscribers, which are informal compared to standard traffic conformance definitions (e.g., those for guaranteed quality of service (QoS) in Internet [19]).

Typically ISPs use traffic shaping to divide the available capacity of a physical access link (e.g., 100+ Mbps by DOCSIS 3.0 [6]) into smaller ones promised to their subscribers per service contracts [20]. With the mechanism like TBF, ISPs regulate the token generation rate, the token bucket size, and optionally the peak rate of combined traffic flows from each subscriber, which provides reasonable QoS to conformant subscribers but prevents non-conformant subscribers from hogging the available bandwidth. At the same time, ISPs want to allow efficient sharing of unused capacity among active subscribers to improve their experience of Internet access, which, like the PowerBoost, is a way to differentiate their access services from their competitors [1].

2.2. Major Issues

The current practice of ISP traffic shaping incurs the following major issues due to its application to the combined traffic flows from a subscriber and the lack of formal definition of traffic conformance as we discussed.

2.2.1. Service Differentiation

Under the current practice of ISP traffic shaping, it is difficult to provide different levels of QoS to different types of traffic flows. For instance, when there are delay-sensitive flows (e.g., VoIP calls) and large non-real-time data flows (e.g., file transfer) from the same subscriber, the current ISP traffic shaping cannot differentiate the former from the latter because it is done per subscriber over combined flows. If traffic shaping is done per individual flow as in integrated services (IntServ) [21] or at least per class as in differentiated services (DiffServ) [22], this issue can be addressed. To do that, however, we need per-flow or per-class service contracts, which is not the case currently.

As workarounds, two traffic control schemes for large bulk data flows under the PowerBoost — one based on intermittent transmission with periodic *on* and *off* cycles and the other based on WonderShaper — are investigated in [2], which significantly improve the latency of delay-sensitive flows while achieving similar long-term rates because these schemes do not deplete the tokens at any time and thereby remove the chance of queueing. The main drawback is that the shaping parameters need to be known in order to exploit this behavior.

2.2.2. Conflicting Requirements

Another major issue is that guaranteeing QoS to the subscribers and enabling efficient sharing of unused capacity among them seemingly contradict each other. For better QoS guarantee, tight burst control is preferred for stricter control of user traffic; for efficient sharing of unused capacity among active subscribers, on the other hand, loose burst control is desirable as in the PowerBoost. Because the average and the peak rates of a service are determined by a service contract and underlying access technology respectively, determining a proper size of the token bucket is a key to ISP traffic shaping.

The effect of ISP traffic shaping, especially the effect of the token bucket size, has been studied with the PowerBoost: In [1], a qualitative investigation of the effect of PowerBoost on TCP and applications is done, while its impact on ISP speed measurements is studied based on the actual results from Sam-Knows measurements. In [2], the effect of the PowerBoost is also studied based on the results obtained from two independent gateway deployments with focus on packet-level performance at the network/transport layers like packet latency and TCP throughput. Even though both studies are based on the measurements from field-deployed home gateways and thereby provide meaningful snapshots of actual broadband access performance, there is

neither systematic investigation on the effect of ISP traffic shaping on user-perceived performance nor consistent conclusion made even for packet-level performance because so many conditions, including the way and the time of measurements and background traffic from other users, are simply out of control in such large-scale field tests. They also lack the investigation of the interaction of shaped traffic flows from multiple subscribers in a common shared access network.

3. Methodology

The shift from packet-level performance measures to user-perceived ones, together with user-behavior-based traffic generation, demands a new methodology for experiments and the analysis of their results. In this section we describe the details of experimental platform and system models, generation of traffic and gathering of performance measures, and a framework for a comparative analysis of the results.

3.1. *Experimental Platform and System Models*

Due to the complexity of protocols and the interactive nature of traffic in the study of network architectures and protocols, researchers now heavily depend on experiments with simulation and/or test beds implementing proposed architectures and protocols rather than mathematical analyses under simplifying assumptions. Especially the experimental platform for this study should be able to capture the interaction of traffic flows through a complete protocol stack, which are generated based on user behavior models at the application/session layers. We do also need a full control of the whole end-to-end network configuration to eliminate the effect of complicated factors on the performance measures of interest (e.g., background traffic in metro and backbone networks). For these reasons, we implemented a virtual test bed composed of detailed simulation models based on OMNeT++ [23] and INET framework [24], which provide models for end-user applications as well as a complete TCP/IP protocol stack.

Fig. 1 shows an overview of the virtual test bed for a shared access network. Virtual local area network (VLAN)-based implementations of the access switch and the subscriber unit are shown in Fig. 2, which abstract key features essential for this study from specific systems like the cable modem termination system (CMTS) and the cable modem for cable Internet and the optical line termination (OLT) and the optical network unit (ONU) for

passive optical networks (PONs). As for the (optical) distribution network ((O)DN), we use a VLAN-aware Ethernet switch to model it. Fig. 3 also shows a model for an end-user which is connected to the subscriber unit through the UNI. Note that there could be multiple *users* who share the broadband connection of a *subscriber* as shown in Fig. 1.⁵

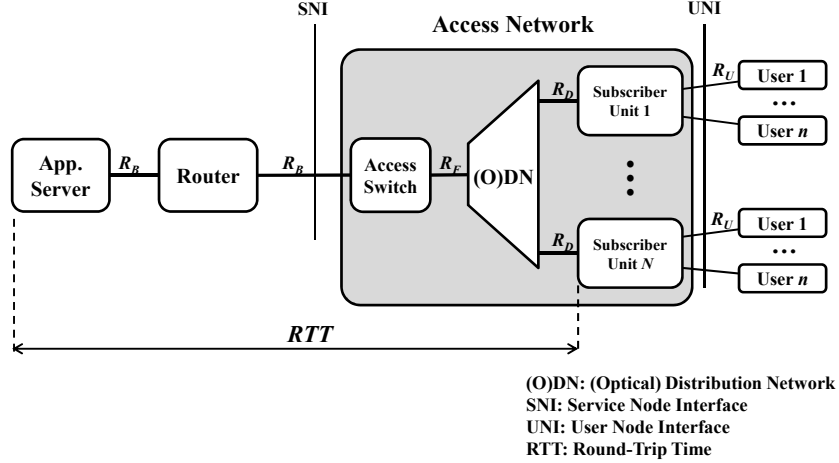


Figure 1: An overview of a virtual test bed for a shared access network.

The reason we adopt VLAN-based abstract models for the access switch, the shared distribution network (i.e., ODN), and the subscriber units is that we need models which can provide features common to specific systems (e.g., cable Internet and Ethernet PON (EPON)), while being practical enough to be compatible with other components and systems of the virtual test bed like the backbone router and the application server implementing standard protocols (e.g., hyper text transfer protocol (HTTP) and file transfer protocol (FTP) over TCP/IP). To identify each subscriber in our shared access model, we assign a unique VLAN identifier (VID) to each subscriber, which is similar to the service identifier (SID) in cable Internet and the logical link identifier (LLID) in EPON. The egress classification in the access switch is based on VIDs and the classified downstream flows go through TBFs and are scheduled by a round-robin scheduler as shown in Fig. 2. Note that, because we mainly focus on the performance of downstream traffic in this paper, we do not

⁵Consider, for example, a household (i.e., a subscriber) where family members (i.e., users) share an Internet connection through a home network.

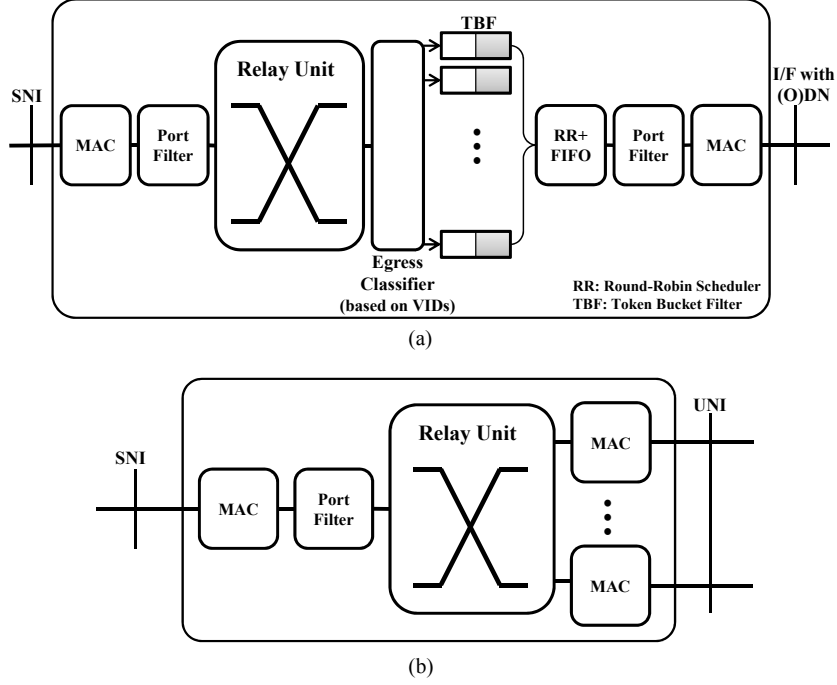


Figure 2: VLAN-based implementations of (a) the access switch and (b) the subscriber unit.

consider upstream traffic shaping at the subscriber unit.

3.2. Traffic Generation and Performance Measurement

As for HTTP and FTP traffic, we use the user-behavior-based traffic model shown in Fig. 4, which is based on the model introduced by the 3rd generation partnership project (3GPP) for CDMA2000 evaluation [25]. The FTP traffic is a special case of this model, where there are a request and response(s) for the main object (e.g., a file to download) only. The parameter values used for experiments are summarized in Table 1. Note that these traffic models and parameter values are gaining wider acceptance among other standard bodies (e.g., WiMAX Forum [26] and IEEE 802.20 [27]) and now serving as reasonable consensus models bringing some uniformity in comparisons of systems.

As metrics of user-perceived performance for HTTP and FTP traffic, we collect packet-call-level performance measures during an experiment. For instance, the web page delay suggested as the main performance metric for

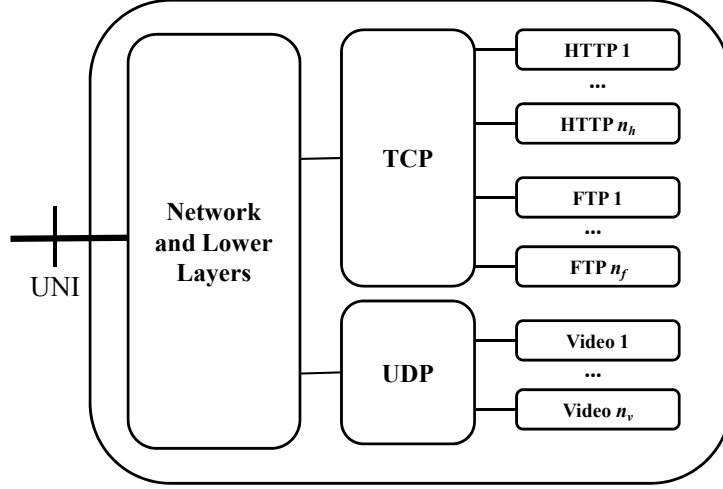


Figure 3: An end-user model.

web browsing in [28] corresponds to the packet call delay defined as the time taken from the beginning to the end of a packet call in Fig. 4. Likewise, the average page throughput and the mean page transfer rate in [28] are defined as the ratio of the mean packet call size (i.e., the size of all objects in a packet call) to the mean packet call delay and the mean of all the packet call size to packet call delay ratios, respectively. Later in the comparative analysis, even though we obtain all three packet call-level metrics for both traffic types, we mainly use the average packet call delay (i.e., the web page delay) and the average packet call throughput as main metrics for HTTP and FTP traffic, respectively.

As for streaming video traffic, we use two video traces for VBR-coded H.264/advanced video coding (AVC) clips from Arizona state university (ASU) video trace library [29] — i.e., the common intermediate format (CIF) “Star Wars IV” and the high definition (HD) format “Terminator2” — whose properties are summarized in Table 2. Frames are encapsulated by real-time transport protocol (RTP) and then user datagram protocol (UDP), and finally carried as the payload of IP packets. The starting frame is selected randomly from the trace at the beginning of simulation and the whole trace is cycled throughout the period of the stream. Once all the frames of a given trace have been processed, the same video is immediately started up, but with a new randomly-selected starting frame. In this way the resulting traffic is kept random without any fixed phase relationship among multiple video

Table 1: Parameter values for HTTP and FTP traffic models [25]

Parameters/Measurements	Best Fit (Parameters)
HTTP Model	
HTML Object Size [Byte]: Mean=10710, SD=25032, Min=100, Max=2M	Truncated Lognormal: $\mu=8.35$, $\sigma=1.37$, Min=100, Max=2M
Embedded Object Size [Byte]: Mean=7758, SD=126168, Min=50, Max=2M	Truncated Lognormal: $\mu=6.17$, $\sigma=2.36$, Min=50, Max=2M
Number of Embedded Objects: Mean=5.64, Max=53	Truncated Pareto ¹ : $\alpha=1.1$, $k=2$, $m=55$
Parsing Time [sec]: Mean=0.13	Exponential: $\lambda=7.69$
Reading Time [sec]: Mean=30	Exponential: $\lambda=0.033$
Request Size [Byte]: Mean=318.59, SD=179.46	Uniform: $a=0$, $b=700$
FTP Model	
File Size [Byte]: Mean=2M, SD=0.722M, Max=5M	Truncated Lognormal: $\mu=14.45$, $\sigma=0.35$, Max=5M
Reading Time [sec]: Mean=180	Exponential: $\lambda=0.006$
Request Size [Byte]: Mean=318.59, SD=179.46	Uniform: $a=0$, $b=700$

¹ k is subtracted from the generated random value to obtain a distribution for the number of embedded objects.

effect of a 5-second de-jitter buffer as suggested in [25]: Because we know a cumulative display time of each frame with respect to the first startup I frame thanks to the decoding frame number in the video trace, we can convert frame delay into frame loss as shown in Algorithm 1.⁶

Initialization;

$i \leftarrow$ decoding frame number of the startup I frame;

$t_i \leftarrow$ arrival time of the startup I frame;

$T_F \leftarrow$ frame period; /* e.g., 33.3 ms */

$T_D \leftarrow$ startup delay; /* e.g., 5 s */

Arrival on the arrival of a video frame with a decoding frame number j and arrival time t_j ;

if $t_j > t_i + T_F \times (j - i) + T_D$ **then**

Discard the arrived frame;

end

Algorithm 1: Frame delay-loss conversion in the video streaming model.

We can also consider an integrated traffic model where both FTP and streaming video traffic are embedded within HTTP traffic flows once *behavioral traffic models* for these cases (especially for embedded streaming video like YouTube) are available.

For details of the implemented traffic models, readers are referred to [33].⁷

3.3. Comparative Analysis Framework

Because our investigation depends on simulation experiments rather than mathematical analyses as discussed in Sec. 3.1, we need a way to systematically take into account the statistical variability in measured data from the experiments. For a comparative analysis of the effect of ISP traffic shaping with respect to the unshaped case, we do also need to collectively process the multiple user-perceived performance metrics for different service types (i.e., HTTP, FTP, and streaming video) of possibly multiple users belonging to

⁶We use this simple model of decoding and play-out buffering; detailed modeling like adaptive media playout scheme in [32] is beyond the scope of this paper.

⁷Note that the simulation models, configurations, and scripts for pre- and post-processing are available online at “<http://github.com/kyeongsoo/inet-hnrl>”.

the same subscriber; note that, while the metrics discussed in Sec. 3.2 are to capture the performance of individual traffic flows perceived by a user, the ISP traffic shaping is done at the subscriber level over the combined traffic flows from the multiple users within the subscriber.

To take into account the statistical variability in measured data and process multiple performance metrics in an integrated way during the comparison, we adopt the comparative analysis framework that we proposed for the clean-slate design of next-generation optical access in [5]. In this framework we use the user-perceived performance of a single subscriber under no traffic shaping as a reference case against which we compare the user-perceived performance of all other shaped configurations either with a single subscriber or with multiple subscribers. Then, using the multivariate non-inferiority testing procedure shown in Fig. 5, we find system configurations (e.g., the number of subscribers and the number of traffic flows per subscriber) and TBF parameter values (e.g., token bucket size and token generation rate) for which the user-perceived performance are statistically *non-inferior* to those of the reference case.

The comparative analysis based on user-perceived performance with respect to a reference case is inspired by the equivalent circuit rate (ECR) measure for a quantitative comparison of hybrid fiber coaxial (HFC) cable-based shared access network and DSL-based dedicated access network architectures in [28]. The original comparison framework for the ECR assumes that each *active* user is using a single application, i.e., web browsing, and bases the comparison on the user-perceived performance of that application only, where there is no differentiation between the user and the subscriber and no statistical comparison procedure is used to take into account the statistical variability in measured data.

The comparative analysis framework proposed in [5] addresses these drawbacks of the original ECR framework by extending it to multiple performance metrics and taking into account statistical variability of measured data in comparison using multivariate non-inferiority testing. The non-inferiority testing is a one-sided variant of the equivalence testing used in Medicine and Biology for the establishment of the equivalence between two different clinical trials or drugs [34]. The non-inferiority testing procedure is based on statistical hypothesis testing and as such takes into account the statistical variability in measured data. To compare multiple performance metrics in an integrated way, the non-inferiority testing is extended by intersection-union testing (IUT) [34]. In this way we can formally define the equivalence of the

results from two configurations (i.e., user-perceived performance in our case).

In the multivariate non-inferiority testing shown in Fig. 5, the null and the alternative hypotheses of the non-inferiority testing for each measure M_i (e.g., web page delay), $i=1, 2, \dots, N_M$, are given by

$$\begin{cases} H_0 : \mu_{i,C} - \mu_{i,R} \geq \delta_i \\ H_1 : \mu_{i,C} - \mu_{i,R} < \delta_i \end{cases} \quad (1)$$

where $\mu_{i,C}$ and $\mu_{i,R}$ denote population means of M_i for the candidate (i.e., shaped) and the reference (i.e., unshaped) configurations, respectively, and δ_i represents the tolerance for the measure M_i . The null hypothesis (H_0) is rejected if the limit of one-sided confidence interval for the difference (i.e., $\mu_{i,C} - \mu_{i,R}$) is less than the tolerance [35]. This means that the result from the candidate configuration is “at least as good as” the reference one for the given measure M_i . Note that for each measure M_i , we need to determine an appropriate tolerance value (δ_i) and, if needed, change the direction of inequalities accordingly. For example, we need to change the hypotheses for the packet call throughput of FTP traffic (unlike delay, higher throughput is better) as follows:

$$\begin{cases} H_0 : \mu_{i,C} - \mu_{i,R} \leq -\delta_i \\ H_1 : \mu_{i,C} - \mu_{i,R} > -\delta_i \end{cases} \quad (2)$$

For details of the comparative analysis framework, readers are referred to [5].

4. Simulation Results

The experiment configurations considered in this paper — i.e., two unshaped (i.e., U_1 and U_2) and eighteen shaped (i.e., $S_{1,1}$ – $S_{1,9}$ and $S_{2,1}$ – $S_{2,9}$) ones — are summarized in Table 3: For all the configurations, the backbone line rate (R_B) and the end-to-end round-trip time (RTT) are fixed to 100 Gbit/s and 10 ms, respectively. As for the access line rate for the distribution (R_D) and the feeder (R_F), two values of 100 Mbit/s and 1 Gbit/s are considered, which could represent the capacities provided by the cable Internet (i.e., DOCSIS 3.0) and the EPON, respectively. Unless stated otherwise, the number of HTTP streams (n_h), the number of FTP streams (n_f), and the number of video streams (n_v) per user are set to 1.

As for the token generation rate (i.e., the long-term average of the service rate per subscriber), we consider the values of 2 Mbit/s, 10 Mbit/s, and 20

Table 3: Summary of experiment configurations

Config.	Network Parameters			TBF Parameters	
	RTT	R_B	R_F, R_D, R_U	TGR ¹	TBS ²
U ₁	10 ms	100 Gbit/s	100 Mbit/s	Unshaped	
S _{1,1}				2 Mbit/s	1 MB ³
S _{1,2}					10 MB
S _{1,3}					100 MB
S _{1,4}				10 Mbit/s	1 MB
S _{1,5}					10 MB
S _{1,6}					100 MB
S _{1,7}				20 Mbit/s	1 MB
S _{1,8}					10 MB
S _{1,9}					100 MB
U ₂			1 Gbit/s	Unshaped	
S _{2,1}				30 Mbit/s	10 MB
S _{2,2}					100 MB
S _{2,3}					1 GB ⁴
S _{2,4}				60 Mbit/s	10 MB
S _{2,5}					100 MB
S _{2,6}					1 GB
S _{2,7}				90 Mbit/s	10 MB
S _{2,8}					100 MB
S _{2,9}					1 GB

¹ Token generation rate.

² Token bucket size.

³ 1 MB = 10⁶ bytes.

⁴ 1 GB = 10⁹ bytes.

Mbit/s for the access line rate of 100 Mbit/s and the values of 30 Mbit/s, 60 Mbit/s, and 90 Mbit/s for the access line rate of 1 Gbit/s.⁸; the minimum 2-Mbit/s service rate is chosen especially because it is a target rate for the *Universal Service Broadband Commitment* in the Digital Britain Final Report [37]. Note that the combined traffic generation rate of the three flows for HTTP, FTP, and streaming video traffic from a single user, which is measured at the physical layer without traffic shaping during preliminary simulations, is 1.83 Mbit/s for configurations U_1 and $S_{1,1}$ – $S_{1,9}$ with the “Star Wars IV” clip and 30 Mbit/s for configurations U_2 and $S_{2,1}$ – $S_{2,9}$ with the “Terminator2” clip. As for the token bucket size, we consider three values of 1 MB, 10 MB, and 100 MB for the access line rate of 100 Mbit/s and ten times those values for the access line rate of 1 Gbit/s. The peak rate of TBF is set to the access line rate except for the cases of investigating its effect discussed in Sec. 4.1.

During the comparative analysis between the unshaped and the shaped configurations, we fix the values of the network parameters and the number of users per subscriber (i.e., n) for both configurations, while we vary the values of TBF parameters and the number of subscribers (i.e., N) for the shaped configurations to investigate the effect of ISP traffic shaping at the subscriber and the access levels.

Each simulation is run for 3 hours with a warmup period of 20 minutes, both in simulation time. To calculate confidence intervals and obtain test statistics for the multivariate non-inferiority testing, each simulation run is repeated ten times with different random number seeds.

4.1. With a Single Subscriber

We first investigate the effect of ISP traffic shaping on the user-perceived performance at the subscriber level with a single subscriber. The amount of incoming traffic to the TBF is controlled by the number of users per subscriber n . During the investigation the major focus is put on token bucket sizes which, for a given token generation rate, can provide user-perceived performance non-inferior to those of a subscriber under no traffic shaping.

Figs. 6 and 7 show representative metrics of user-perceived performance for a single subscriber, where we observe that the effect of token bucket size is

⁸For the values of token generation rates, we referred to the current Virgin Media Cable traffic management policy [36].

prominent for both HTTP and the FTP traffic; as for the DFR of streaming video, the effect of token bucket size is negligible for all the configurations with the access line rate of 100 Mbit/s (except for token bucket size of 1 MB), while it is rather significant with the access line rate of 1 Gbit/s where the ratio of video traffic to the total traffic is significantly higher. The large effect of token bucket size on file transfer performance is what we can expect from the discussions in [1], but we found out that the effect of token bucket size on the average HTTP page delay is also quite significant as the combined traffic generation rate approaches to the token generation rate (e.g., $n=1$ for token generation rate of 2 Mbit/s, $n=5$ for 10 Mbit/s, and $n=10$ for 20 Mbit/s); even in such a condition, however, we also note that the large token bucket size — i.e., 100 MB and 1 GB for access line rates of 100 Mbit/s and 1 Gbit/s, respectively — can provide user-perceived performance comparable to those without traffic shaping.

For a comparative analysis, we carried out the multivariate non-inferiority testing described in Sec. 3.3 for the shaped configurations (i.e., $S_{1,1}$ – $S_{1,9}$ and $S_{2,1}$ – $S_{2,9}$) with respect to the corresponding reference configurations (i.e., U_1 and U_2). We set the tolerance (i.e., δ_i) to 10 percent of the sample mean of performance measure for the reference case and the significance level — i.e., the probability of rejecting the null hypothesis H_0 when it is true [38, Section 8.1.2] — to 0.05. The results are shown in Fig. 8, where $\max(n_{eqv})$ is defined as the maximum number of users per subscriber (i.e., n) of a shaped configuration which provides user-perceived performance non-inferior to those of the corresponding unshaped configuration given the same number of users per subscriber and the access line rate.

The results for the access line rate of 100 Mbit/s in Fig. 8 (a) show that for the token generation rate of 2 Mbit/s, the token bucket size of 100 MB can support one user with user-perceived performance non-inferior to those without traffic shaping, while for the token generation rates of 10 Mbit/s and 20 Mbit/s, the same token bucket size can support up to five and ten users respectively. Note that the traffic from one, five, and ten users per subscriber fully loads the TBF with the token generation rates of 2 Mbit/s, 10 Mbit/s, and 20 Mbit/s. It is remarkable to see that with the token bucket size of 100 MB, the token generation rate of mere 2 Mbit/s can provide user-perceived performance nearly equivalent to those with the access line rate of 100 Mbit/s, the rate fifty times higher than the token generation rate. Similar observations are made for the results for the access line rate of 1 Gbit/s in Fig. 8 (b).

With two configurations $S_{1,3}$ (i.e., token generation rate of 2 Mbit/s and token bucket size of 100 MB for access line rate of 100 Mbit/s) and $S_{2,3}$ (i.e., token generation rate of 30 Mbit/s and token bucket size of 1 GB for access line rate of 1 Gbit/s) which can support up to one user per subscriber with user-perceived performance non-inferior to those of unshaped configuration U_1 and U_2 respectively, we investigated the effect of the peak rate on user-perceived performance as shown in Fig. 9, where we vary the peak rate from 2 Mbit/s to 100 Mbit/s for $S_{1,3}$ and from 30 Mbit/s to 1 Gbit/s for $S_{2,3}$. The results show that the peak rate has a significant impact on the user-perceived performance until it increases to five times the average rate for both line rates. As for the streaming video, we found out that, when the peak rate is reduced to the average rate, the resulting decrease in performance is more significant for $S_{2,3}$ than $S_{1,3}$ because the combined rate of traffic flows for $S_{2,3}$ is more close to the average rate than that of $S_{1,3}$. We carried out the investigation of the effect of the peak rate with other configurations as well and observed similar results.

4.2. With Multiple Subscribers

Based on the results of the investigation with a single subscriber, now we study the interaction of shaped traffic flows from multiple subscribers who share the capacity of a common feeder link (i.e., R_F) in a shared access network and their impact on user-perceived performance. The results in Sec. 4.1 suggest that increasing the token bucket size (e.g., from 1 MB to 10 MB to 100 MB for the access line rate of 100 Mbit/s) improves user-perceived performance at the subscriber level, which is a good news from end-users' perspective. The large token bucket size and the resulting large bursts from each subscriber's traffic, on the other hand, may have negative impacts on the user-perceived performance at the access level due to their interaction on the common shared link. From ISPs' point of view, it is interesting to see how many subscribers, given the access line rate and the TBF parameters, can be supported with user-perceived performance non-inferior to those of a corresponding unshaped, single-subscriber configuration. Note that it is well known that a large token bucket size increases the deterministic end-to-end packet delay bounds of various work-conserving scheduling disciplines with TBF-constrained traffic [17].

The results of the multivariate non-inferiority testing for shaped configurations with respect to the reference cases (e.g., $S_{1,5}$ and $S_{1,6}$ with $n=4$ with respect to U_1 with $n=4$) are summarized in Table 4, where $\max(N_{eqv})$

Table 4: Results of multivariate non-inferiority testing

Config.	n	$\max(N_{eqv})$	$n \cdot \max(N_{eqv})^1$
S _{1,3}	1	36	36
S _{1,4}	3	13	39
S _{1,5}	3	15	45
	4	11	44
S _{1,6}	3	15	45
	4	11	44
	5	9	45
S _{1,7}	7	6	42
S _{1,8}	7	6	42
	9	5	45
S _{1,9}	7	6	42
	9	5	45
	10	4	40
S _{2,3}	1	30	30
S _{2,4}	1	27	27
S _{2,5}	1	27	27
S _{2,6}	1	27	27
	2	14	28
S _{2,7}	2	14	28
S _{2,8}	2	14	28
S _{2,9}	2	14	28
	3	9	27

¹ Total number of users in the shared access that can be supported with user-perceived performance non-inferior to those of unshaped, dedicated access.

is defined as the maximum number of subscribers that can be supported with user-perceived performance non-inferior to those with the unshaped, single-subscriber configuration with the same number of users per subscriber. The conditions for the multivariate non-inferiority testing are the same as in Sec. 4.1.

The results in Table 4 show that, unlike its impact on packet-level performance, increasing the token bucket size (given the configurations) does not have a negative impact on the user-perceived performance at the access level; given the average rate and the number of users per subscriber, the configurations with larger token bucket size can support as many subscribers as those with smaller token bucket size (e.g., $S_{1,5}$ vs. $S_{1,6}$ for $n=3, 4$) or even more (i.e., $S_{1,4}$ vs. $S_{1,5}$ and $S_{1,6}$ for $n=3$). Considering that we can support more users per subscriber (e.g., up to 3 users for $S_{1,4}$ vs. up to 5 users for $S_{1,6}$) with a larger token bucket size, we can also increase the total number of users — i.e., $n \cdot \max(N_{eqv})$ — by proper dimensioning of TBF parameters in the access network which can be supported with user-perceived performance non-inferior to those with the unshaped, single-subscriber configuration with the same number of users per subscriber.

These results suggest that the large token bucket size in ISP traffic shaping could improve user-perceived performance of each subscriber at both subscriber and access levels, which would be a good news not only for end-users but also for ISPs. Our investigation in this paper, however, does not consider the potential negative impact of the loose burst control on metro and backbone networks, where, unlike access networks, packet-level performance measures are still important; note that we set the backbone line rate (i.e., R_B) to a value much higher than the rate of combined traffic flows from all the subscribers in the access network to prevent it from being a bottleneck during the experiments.

4.2.1. With Non-Conformant Subscribers

We also investigated the effect of loose burst control in the presence of non-conformant subscribers. The experiment configuration for this investigation is shown in Fig. 10, where there are two groups of subscribers, i.e., Group 1 for 10 conformant subscribers with $n=3$ and Group 2 for non-conformant ones with $n=5$. In this experiment we set n_h , n_f and n_v to 1 for the users in Group 1, while we vary n_f with n_h and n_v fixed to 1 for the users in Group 2 to change network load. As for a token bucket, we consider the size of up to 10 GB, which is about a thousand times larger than those provided by cable broadband companies through PowerBoost/Speedboost technologies [1].

Fig. 11 shows that the loose burst control resulting from the large token bucket size up to 1 GB does not negatively affect user-perceived performance with multiple subscribers even in the presence of non-conformant subscribers;

with a much larger token bucket size of 10 GB, however, the negative effect of non-conformant subscribers on the user-perceived performance of conformant subscribers becomes finally visible because the maximum burst duration in this case, which is larger than average inter-session gaps of traffic models (e.g., 30 and 180 seconds for HTTP and FTP services), makes the impact of token bucket size and that of token generation rate virtually indistinguishable.

Fig. 12 shows the results for the same configuration of Fig. 10 but with first-in, first-out (FIFO) scheduling instead of round-robin, where the overall performance becomes worse in general compared to that of round-robin scheduling. Note that, in case of FIFO scheduling, the negative effect of non-conformant subscribers becomes clear with the token bucket size of 1 GB.

5. Conclusions

In this paper we have investigated the effect of ISP traffic shaping on user-perceived performance based on user-behavior-based traffic models for HTTP and FTP services and a real-trace-based one for streaming video and application/session-layer performance metrics. The results from extensive simulations show that a larger token bucket (i.e., up to 100 MB and 1 GB for 100-Mbit/s and 1-Gbit/s access line rates) provides better user-perceived performance at both subscriber and access levels. This implies that the loose burst control (i.e., allowing users to send their traffic at a peak rate, much higher than the average service rate) enables to exploit well the burstiness of real traffic in different time scales and at multiple layers — i.e., user behaviors (e.g., reading time in web browsing) and VBR encoding at the session layer and TCP congestion control at the transport layer — in statistical multiplexing in the shared access network. Regarding any negative impact of the loose burst control, on the other hand, we do not observe any significant disadvantage with a larger token bucket in terms of the user-perceived performance of conformant subscribers even in the presence of non-conformant subscribers again up to 100 MB for the access line rate of 100 Mbit/s; with a much larger token bucket (e.g., size of 10 GB), however, the negative effect of non-conformant subscribers on the user-perceived performance of conformant subscribers becomes clearly visible because the impact of token bucket size and that of token generation rate are virtually indistinguishable in this case.

The results from the current work can provide ISPs valuable insights into the design, deployment, and operation of the next-generation access networks from end-users’ perspective, especially for the control of peak rate and burstiness to improve user-perceived performance for their access services. There are also implications to researchers in the design of next-generation access architectures and protocols, where we need to study a way to better exploit the burstiness of end-user traffic in different time scales and at multiple layers in statistical multiplexing. Still, we do need more investigations with a wider range of network and traffic configurations to reach a firm conclusion on the effect of ISP traffic shaping on user-perceived performance.

The rather different outcomes from this work, compared to those based on traditional packet-level traffic models and performance measures, clearly show the very importance of user-oriented research framework in the study of access network architectures and protocols as discussed in [7, 5].

One of the major difficulties in this work was the lack of established/standardized behavioral traffic models at higher service and access line rates. For instance, due to the lack of higher-rate FTP traffic model, we have to use multiple FTP traffic streams to increase the load to the system instead of single higher-rate stream for non-conformant subscribers. With standardized sets of traffic models together with performance metrics, we could provide both ISPs and end-users benchmarks and/or rating systems useful for comparison shopping of broadband access services from ISPs. In fact, all the efforts described in this paper are aiming at the creation of new benchmarks and/or rating systems for next-generation access and the adoption of them both by ISPs and end-users for advertising and selecting new service plans. There is already a proposal called “Internet Nutrition Labels” [39] in this regard, but it is still based on traditional network-level performance measures. Our plan is to have new rating systems based on the comparative analysis framework described in Sec. 3.3 under several representative workloads, e.g., user behavior models for web browsing, Internet voice/video calls, multimedia streaming, and online gaming. In this way, the design, deployment, and operation of next-generation access networks will be more energy and cost-efficient by properly managing network resources based on actual user behaviours, not worst-case traffic, and with a direct focus on user-perceived performance.

Another area of research for further work is the extension of the multi-variate non-inferiority testing to quantiles/percentiles, especially for delay, because it is quite challenging to obtain not only the quantiles themselves

but also confidence intervals needed for statistical hypothesis testing [40, 41].

Acknowledgement

This paper was presented in part at FOAN 2012, St. Petersburg, Russia, October 2012. This work was supported in part by Amazon Web Services (AWS) in Education Research Grant.

References

- [1] S. Bauer, D. Clark, W. Lehr, PowerBoost, in: Proc. of HomeNets'11, ACM, New York, NY, USA, 2011, pp. 7–12. doi:<http://doi.acm.org/10.1145/2018567.2018570>.
- [2] S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, A. Pescapè, Broadband Internet performance: A view from the gateway, in: Proc. of SIGCOMM'11, Toronto, Ontario, Canada, 2011, pp. 134–145.
- [3] P. Kanuparth, C. Dovrolis, End-to-end detection of ISP traffic shaping using active and passive methods, Technical Report, Georgia Tech, 2011.
- [4] M. Marcon, M. Dischinger, K. Gummadi, A. Vahdat, The local and global effects of traffic shaping in the Internet, in: Communication Systems and Networks (COMSNETS), 2011 Third International Conference on, 2011, pp. 1–10. doi:[10.1109/COMSNETS.2011.5716420](https://doi.org/10.1109/COMSNETS.2011.5716420).
- [5] K. S. Kim, A research framework for the clean-slate design of next-generation optical access, in: Proc. of ICUMT 2011, Budapest, Hungary, 2011, pp. 1–8.
- [6] CableLabs, Docsis 3.0: MAC and upper layer protocols interface specification, CM-SP-MULPIv3.0, 2012.
- [7] K. S. Kim, K. Ennser, Y. K. Dwivedi, Clean-slate design of next-generation optical access, in: Proc. of the 13th International Conference on Transparent Optical Networks (ICTON) (*invited paper*), Stockholm, Sweden, 2011.
- [8] J. S. Turner, New directions in communications (or which way to the information age?), IEEE Commun. Mag. 24 (1986) 8–15.

- [9] A. I. Elwalid, D. Mitra, Stochastic fluid models in the analysis of access regulation in high speed networks, in: Proc. of GLOBECOM'91, 1991, pp. 1626–1632.
- [10] G. de Veciana, Leaky buckets and optimal self-tuning rate control, in: Proc. of IEEE GLOBECOM'94, 1994, pp. 1207–1211.
- [11] Y. H. Kim, B. C. Shin, C. K. Un, Performance analysis of leaky-bucket bandwidth enforcement strategy for bursty traffics in an ATM networks, *Computer Networks and ISDN Systems* 25 (1992) 295–304.
- [12] N. Yin, M. G. Hluchyj, Analysis of the leaky bucket algorithm for on-off data sources, in: Proc. of GLOBECOM'91, 1991, pp. 254–260.
- [13] R. G. Garroppo, S. Giordano, M. Pagano, Estimation of token bucket parameters for aggregated voip sources, *Int. J. Commun. Syst.* 15 (2002) 851–866.
- [14] G. Procissi, A. Garg, M. Gerla, M. Sanadidi, Token bucket characterization of long-range dependent traffic, *Computer Communications* 25 (2002) 1009–1017.
- [15] S. Chong, S.-Q. Li, Spectral analysis of access rate control in high speed networks, *Lecture Notes in Computer Science* 9 (1996) 237–252.
- [16] S. Bregni, P. Giacomazzi, G. Saddemi, Characterization of long-range dependent traffic regulated by leaky-bucket policers and shapers, *Computer Communications* 33 (2010) 714–720.
- [17] H. Zhang, Service disciplines for guaranteed performance service in packet-switching networks, *Proceedings of the IEEE* 83 (1995) 1374–1396.
- [18] ITU-T, ITU-T Recommendation I.371, traffic control and congestion control in B-ISDN, ITU, 1993.
- [19] S. Shenker, C. Partridge, R. Guerin, Specification of guaranteed quality of service, RFC 2212 (Proposed Standard), 1997.
- [20] K. Lakshminarayanan, V. N. Padmanabhan, J. Padhye, Bandwidth estimation in broadband access networks, in: Proc. of the 4th ACM

- SIGCOMM conference on Internet measurement, IMC '04, ACM, New York, NY, USA, 2004, pp. 314–321. doi:<http://doi.acm.org/10.1145/1028788.1028832>.
- [21] R. Braden, D. Clark, S. Shenker, Integrated services in the Internet architecture: an overview, RFC 1633 (Informational), 1994.
 - [22] S. Blake, D. L. Black, M. A. Carlson, E. Davies, Z. Wang, W. Weiss, An architecture for differentiated services, RFC 2475 (Informational), 1998. Updated by RFC 3260.
 - [23] A. Varga, The OMNeT++ discrete event simulation system, in: Proc. of the European Simulation Multiconference (PESM2001), Prague, Czech Republic, 2001, pp. 319–324. URL: <http://www.omnetpp.org/>.
 - [24] A. Varga, et al., INET framework for OMNeT++ 4.0, URL: <http://inet.omnetpp.org/>.
 - [25] cdma2000 evaluation methodology, 3GPP2 C.R1002-B, 2009.
 - [26] WiMAX system evaluation methodology, 2008.
 - [27] Traffic models for IEEE 802.20 MBWA system simulations, Draft 802.20 Permanent Document, 2003.
 - [28] N. K. Shankaranarayanan, Z. J. P. Mishra, User-perceived performance of web-browsing and interactive data in HFC cable access networks, in: Proc. of ICC'01, volume 4, 2001, pp. 1264–1268.
 - [29] G. V. der Auwera, P. T. David, M. Reisslein, Traffic and quality characterization of single-layer video streams encoded with H.264/AVC advanced video coding standard and scalable video coding extension, IEEE Trans. Broadcast. 54 (2008) 698–718.
 - [30] P. Seeling, M. Reisslein, B. Kulapala, Network performance evaluation using frame size and quality traces of single-layer and two-layer video: A tutorial, IEEE Commun. Surveys Tuts. 6 (2004) 58–78.
 - [31] A. Ziviani, B. E. Wolfinger, J. F. Rezende, O. C. Duarte, S. Fdida, Joint adoption of QoS schemes for MPEG streams, Multimedia Tools Appl. 26 (2005) 59–80.

- [32] M. Kalman, E. Steinbach, B. Girod, Adaptive media playout for low-delay video streaming over error-prone channels, *IEEE Trans. Circuits Syst. Video Technol.* 14 (2004) 841–851.
- [33] K. S. Kim, Integration of OMNeT++ hybrid TDM/WDM-PON models into INET framework, OMNeT++ Workshop 2011 code contribution, 2011. URL: http://www.omnet-workshop.org/2011/uploads/slides/OMNeT_WS2011_S5_C2_Kim.pdf.
- [34] R. L. Berger, J. C. Hsu, Bioequivalence trials, intersection-union tests, and equivalence confidence sets, *Statistical Science* 11 (1996) 283–319.
- [35] G. T. da Silva, B. R. Logan, J. P. Klein, Methods for equivalence and noninferiority testing, *Biology of Blood and Marrow Transplantation* 15 (2009) 120–127.
- [36] Virgin media cable traffic management policy, URL: http://help.virginmedia.com/system/selfservice.controller?CMD=VIEW_ARTICLE&ARTICLE_ID=2781&CURRENT_CMD=SEARCH&CONFIGURATION=1002&PARTITION_ID=1&USERTYPE=1&LANGUAGE=en&COUNTY=us&VM_CUSTOMER_TYPE=Cable.
- [37] Department of Culture, Media and Sport and Department for Business, Innovation and Skills, UK, Digital Britain final report, 2009.
- [38] D. C. Montgomery, G. C. Runger, Applied statistics and probability for engineers, John Wiley & Sons, 1994.
- [39] S. Sundaresan, N. Feamster, R. Teixeira, A. Tang, W. K. Edwards, R. E. Grinter, M. Chetty, W. de Donato, Helping users shop for ISPs with Internet nutrition labels, in: *Proc. of HomeNets’11*, New York, NY, USA, 2011, pp. 13–18.
- [40] J.-S. R. Lee, D. McNicle, K. Pawlikowski, Quantile estimations in sequential steady-state simulation, in: *Proc. of the European Simulation Multiconference (ESM’99)*, Warsaw, Poland, 1999, pp. 168–174.
- [41] M. Eickhoff, D. McNickle, K. Pawlikowski, Using parallel replications for sequential estimation of multiple steady state quantiles, in: *Proc. of the 2nd International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS 2007)*, Nantes, France, 2007.

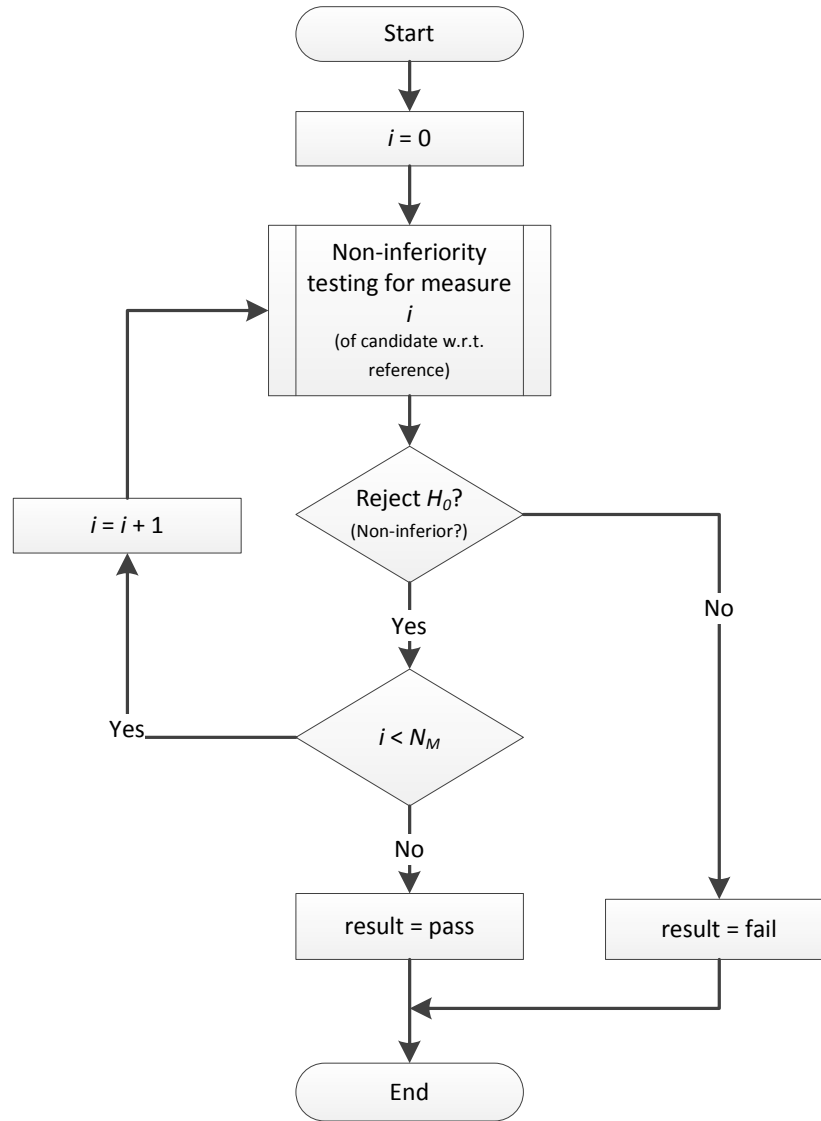
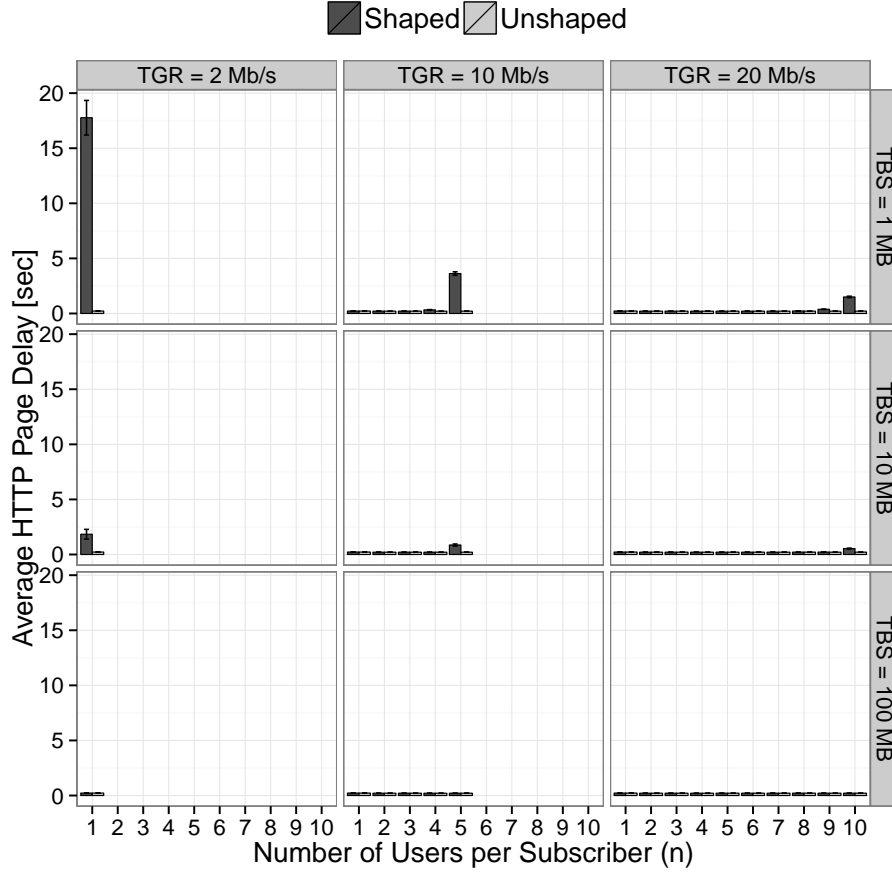
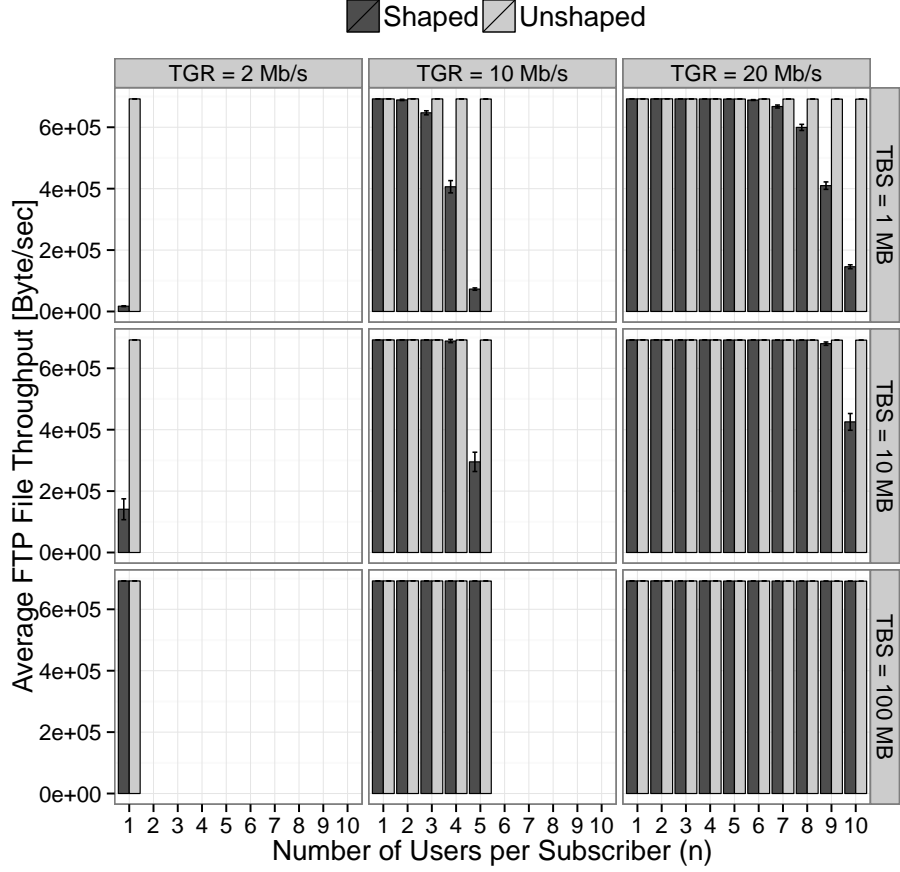


Figure 5: An overview of multivariate non-inferiority testing procedure [5].



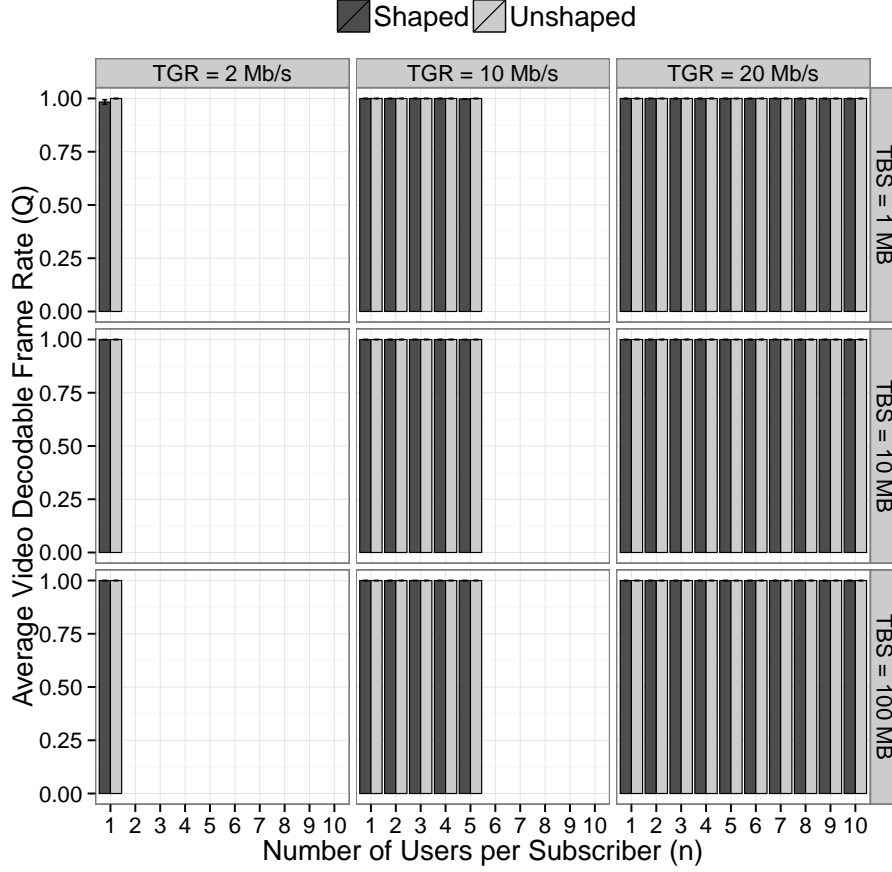
(a)

Figure 6: User-perceived performance metrics with 95 percent confidence intervals for a single subscriber with access line rate of 100 Mbit/s: (a) Average session delay of HTTP traffic, (b) average session throughput of FTP traffic, and (c) decodable frame rate (Q) of UDP streaming video.



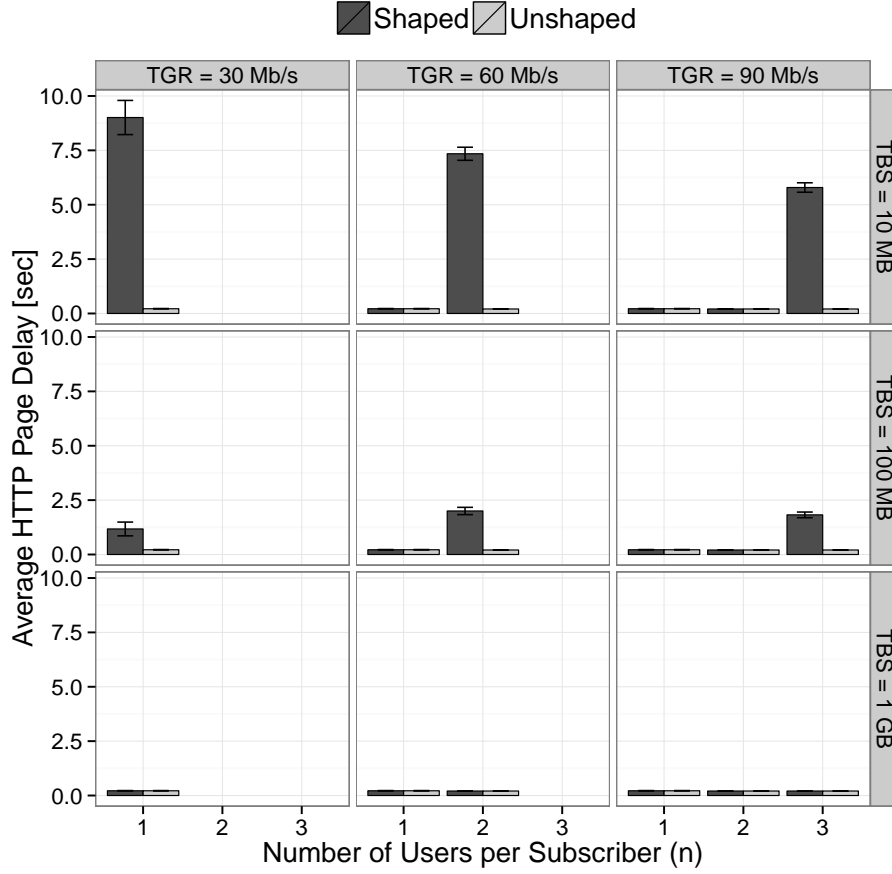
(b)

Figure 6: User-perceived performance metrics with 95 percent confidence intervals for a single subscriber with access line rate of 100 Mbit/s: (a) Average session delay of HTTP traffic, (b) average session throughput of FTP traffic, and (c) decodable frame rate (Q) of UDP streaming video.



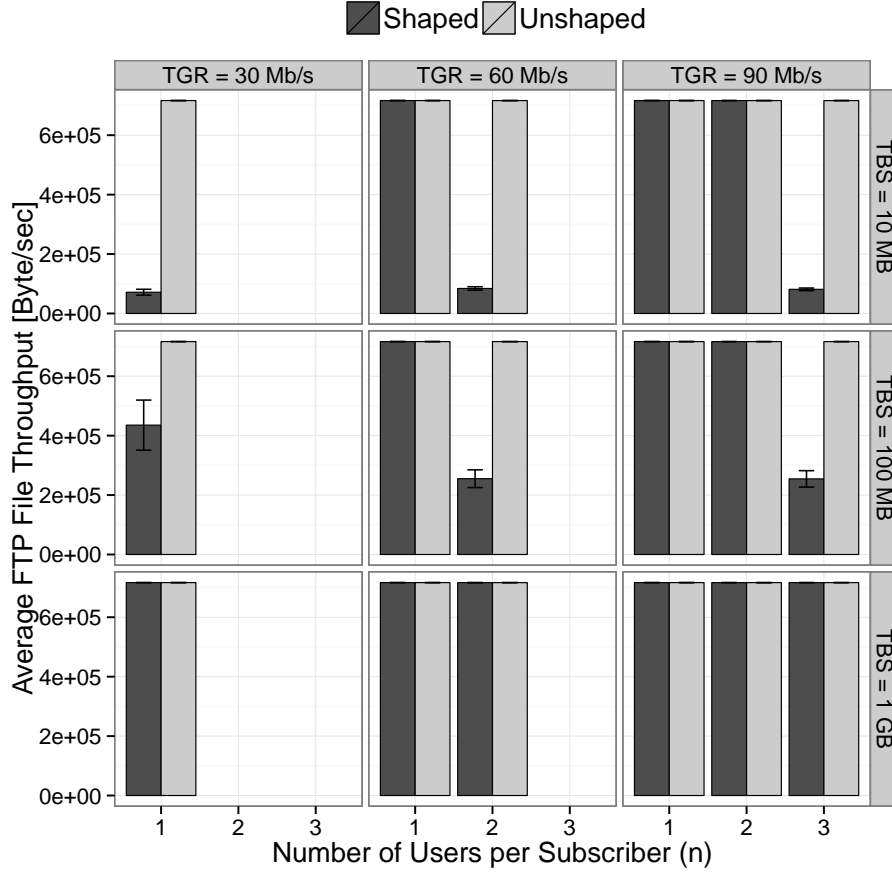
(c)

Figure 6: User-perceived performance metrics with 95 percent confidence intervals for a single subscriber with access line rate of 100 Mbit/s: (a) Average session delay of HTTP traffic, (b) average session throughput of FTP traffic, and (c) decodable frame rate (Q) of UDP streaming video.



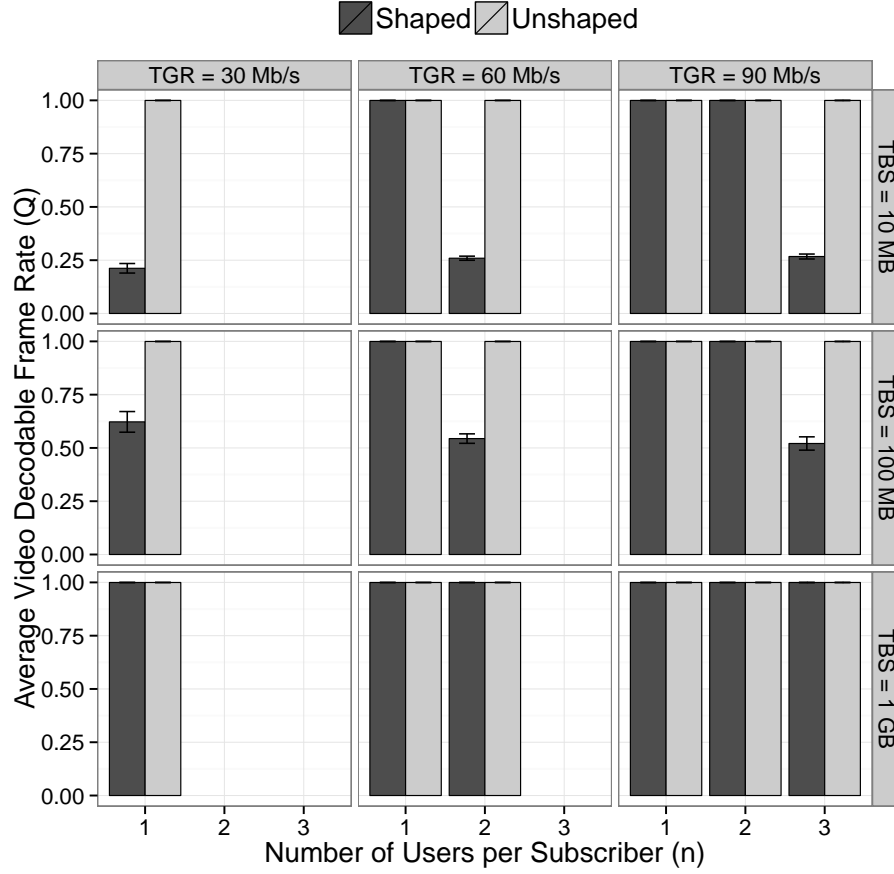
(a)

Figure 7: User-perceived performance metrics with 95 percent confidence intervals for a single subscriber with access line rate of 1 Gbit/s: (a) Average session delay of HTTP traffic, (b) average session throughput of FTP traffic, and (c) decodable frame rate (Q) of UDP streaming video.



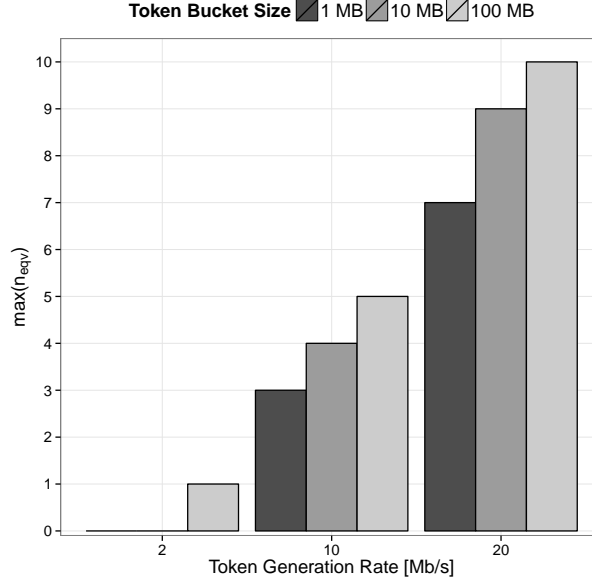
(b)

Figure 7: User-perceived performance metrics with 95 percent confidence intervals for a single subscriber with access line rate of 1 Gbit/s: (a) Average session delay of HTTP traffic, (b) average session throughput of FTP traffic, and (c) decodable frame rate (Q) of UDP streaming video.

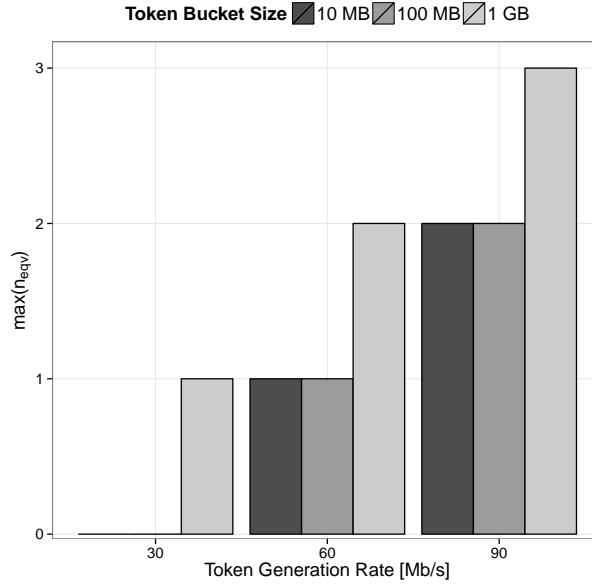


(c)

Figure 7: User-perceived performance metrics with 95 percent confidence intervals for a single subscriber with access line rate of 1 Gbit/s: (a) Average session delay of HTTP traffic, (b) average session throughput of FTP traffic, and (c) decodable frame rate (Q) of UDP streaming video.

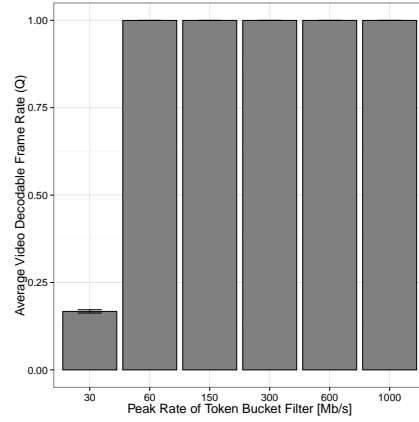
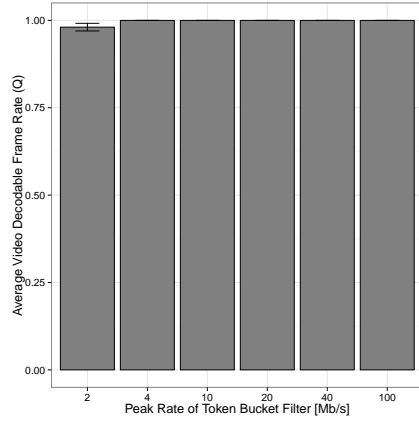
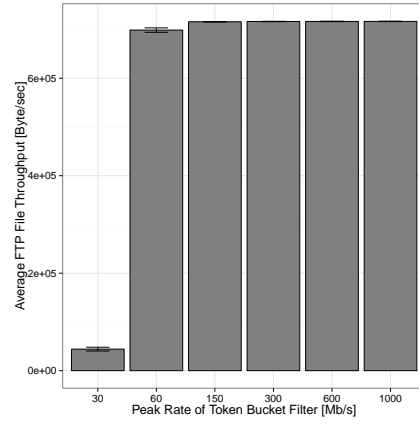
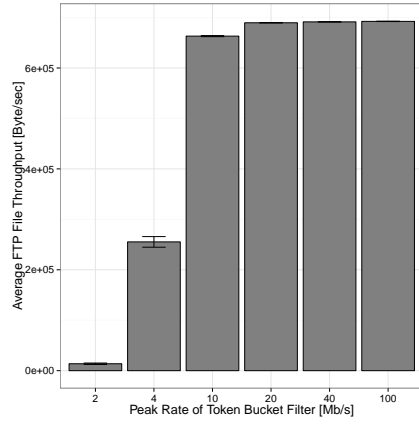
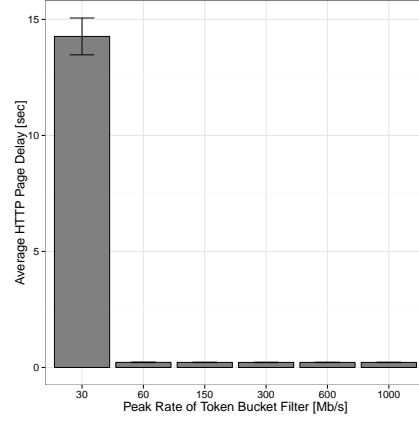
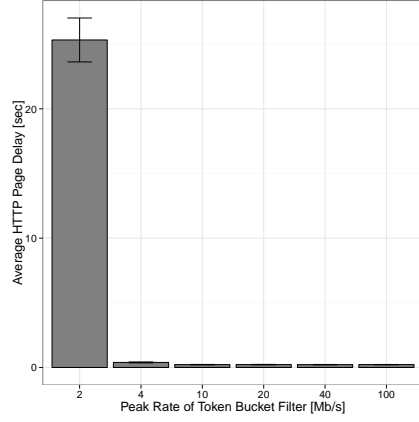


(a)



(b)

Figure 8: Maximum number of users per subscriber of shaped configurations which provides user-perceived performance non-inferior to that of the corresponding unshaped configurations for a single subscriber with access line rate of (a) 100 Mbit/s and (b) 1 Gbit/s.



(a)

(b)

Figure 9: Effect of the peak rate on user-perceived performance with a single subscriber and a single user: (a) access line rate of 100 Mbit/s with token generation rate of 2 Mbit/s and token bucket size of 100 MB and (b) access line rate of 1 Gbit/s with token generation rate of 30 Mbit/s and token bucket size of 1 GB.

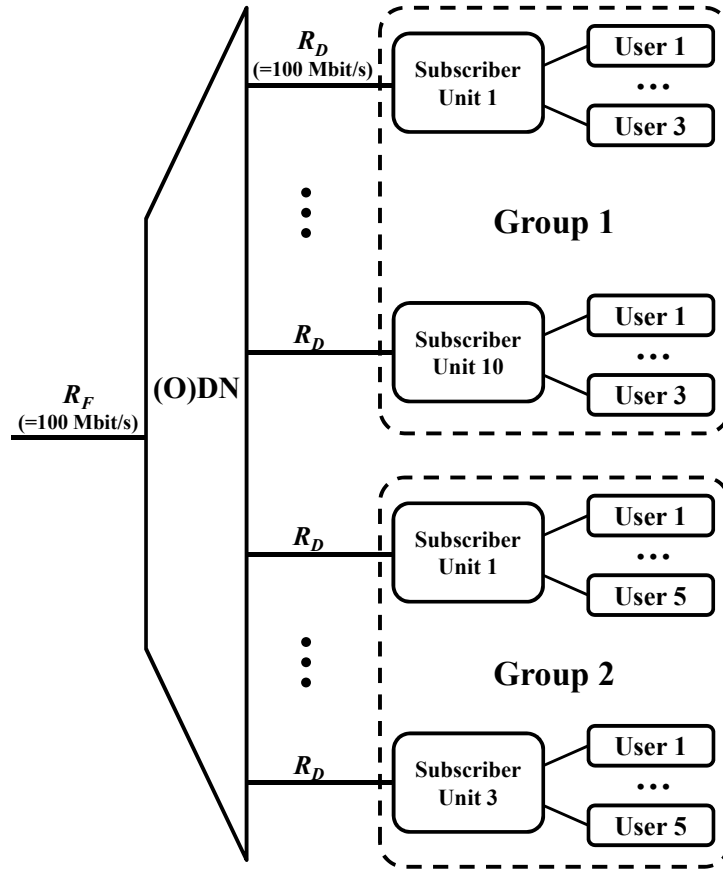
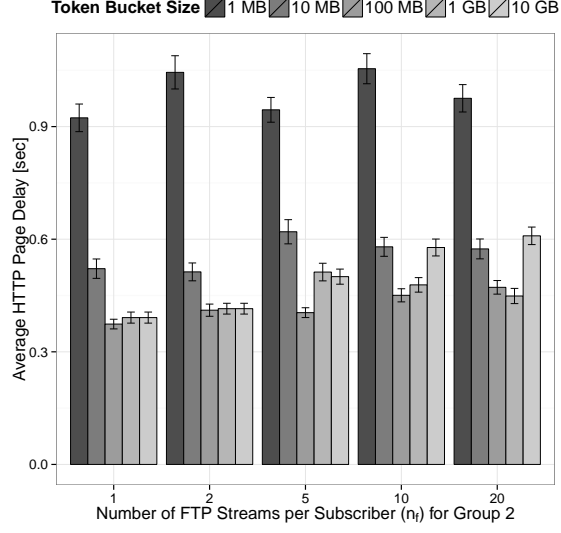
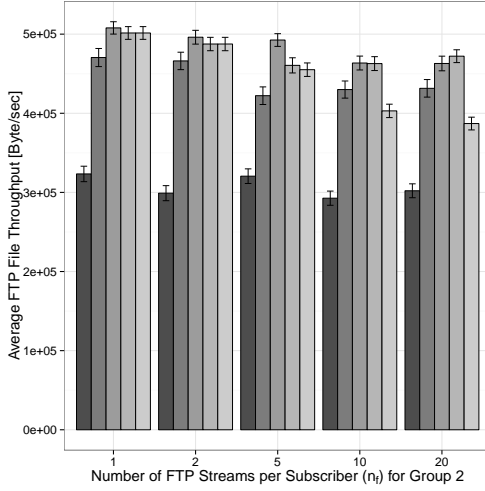


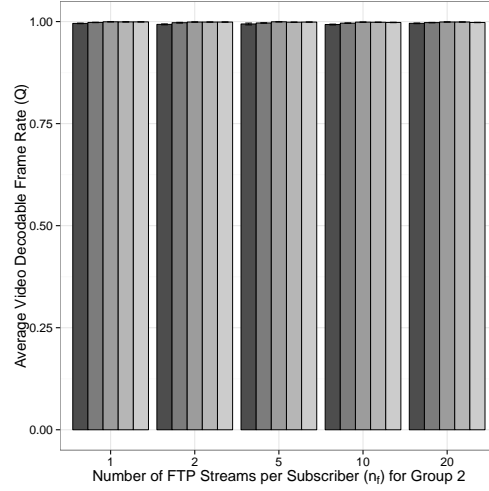
Figure 10: Unbalanced network configuration with two groups of subscribers and access line rate of 100 Mbit/s.



(a)

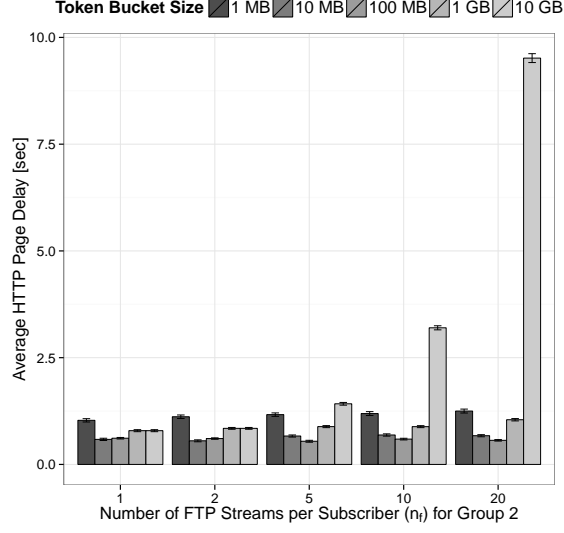


(b)

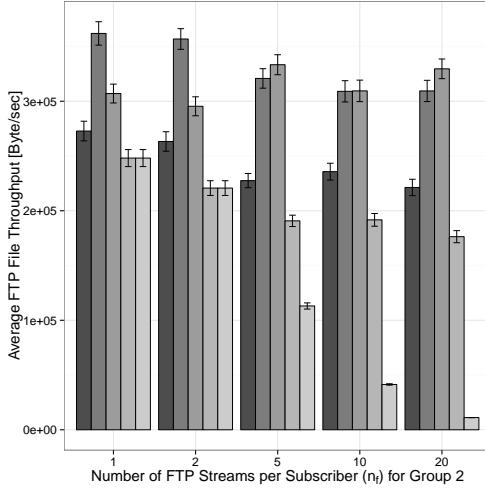


(c)

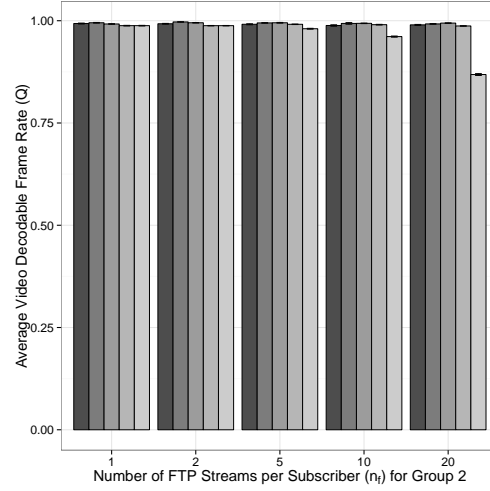
Figure 11: User-perceived performance measures with 95 percent confidence intervals for Group 1 of the network configuration shown in Fig. 10 with token generation rate of 10 Mbit/s: (a) Average session delay of HTTP traffic, (b) average session throughput of FTP traffic, and (c) decodable frame rate (Q) of UDP streaming video.



(a)



(b)



(c)

Figure 12: User-perceived performance measures with 95 percent confidence intervals for Group 1 of the network configuration shown in Fig. 10 with *FIFO scheduling* and token generation rate of 10 Mbit/s: (a) Average session delay of HTTP traffic, (b) average session throughput of FTP traffic, and (c) decodable frame rate (Q) of UDP streaming video.