Available online at www.sciencedirect.com
ScienceDirect
Computer Communications 31 (2008) 2257–2269
computer communications
www.elsevier.com/locate/comcom

Network utility maximization for triple-play services

Lei Shi *, Changbin Liu, Bin Liu

Department of Computer Science and Technology, Tsinghua University, Room 9-416, East Main Building, Beijing 100084, PR China

Received 22 April 2007; received in revised form 14 February 2008; accepted 15 February 2008
Available online 29 February 2008

Abstract

It is well known that Next-Generation Network (NGN) will inevitably carry triple-play services (i.e. voice, video and data) simultaneously. However, the traditional strict-priority based scheduling algorithm intensively used in current Internet cannot maximize the overall network utility for NGN, instead brings significant global welfare loss. In this paper, we study how to achieve Network Utility Maximization (NUM) in NGN running triple-play services. By investigating the characteristics of most of its traffic classes, we explicitly present their utilities as the function of allocated bandwidth. We further formulate the NUM objective as a nonlinear programming problem with both inequality and equality constraints. A solution using Lagrange Multiplier is given on the simplified problem with only equality constraints, which indicates the major distinction from strict-priority based scheduling, the existence of a turning point for IPTV users. Simulations are also carried out using LINGO on the original complicated problem. Several useful results are presented on the new features of the NUM-based scheduling. We also discuss the methods to alleviate the impact of turning point and the consequent unstable bandwidth allocation.

Keywords: Network utility maximization; Triple-play services; Next generation network; Bandwidth allocation

1. Introduction

The Internet has been evolving in recent years to adapt with the emerging abundant applications. Specifically, in the near future, the voice, video and data traffic (herein called triple-play services), which are previously forwarded by separate networks such as Public Switched Telephone Network (PSTN), the cable television network and the original Internet, will be carried on a single converged network, i.e. the Next Generation Network (NGN). NGN must be able to natively conduct triple-play services, which means that all traffic classes of voice, video and data should be managed to meet their particular Quality of Service (QoS) requirements, such as strict packet delay, jitter and loss guarantees. It is believed that the deployment of NGN and the provisioning of triple-play services will eventually not only benefit the Internet users with richer contents, but also increase ISP revenues by acquiring much higher per-subscriber profit.

While several Internet Service Providers (ISPs) have proposed their architecture and detailed specifications to support triple-plays in NGN, they all have to deal with a critical issue that how to schedule traffic and allocate bandwidth for triple-play services at both backbone networks and access links. Due to the efficiency consideration that NGN cannot be designed with over-provisioning technique to avoid congestion, more advanced congestion-phased traffic scheduling algorithms are essentially required to compromise the benefit of all the traffic classes. Designing such a scheduling (bandwidth allocation) algorithm is exactly the premier issue this paper tries to settle.

Prior to the study presented here, numerous related works have been published on this issue. In industry designing NGN [1,3], the strict-priority based scheduling algorithm is the most widely adopted one in carrying out bandwidth allocation. However, this solution rigidly favors the voice and

* Corresponding author. Tel.: +86 10 62773441.
E-mail address: shijim@gmail.com (L. Shi).

0140-3664/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2008.02.016

video traffic without flexibility, thus can only be deemed as a conservative method when no better one is available.

On the other hand, researches in academia mainly concentrated on utility-based solutions. Shenker [16] for the first time discussed traffic classifications in IP network from the viewpoint of user utility. He further investigated the characteristics of several well-known traffic classes, including TCP elastic traffic, UDP hard real-time traffic, delay-adaptive traffic, as well as rate-adaptive traffic. However, no explicit expression of utility function was given. Later, Kelly et al. presented some first works [11,12] in applying utility-based method from economics to the area of scheduling and bandwidth allocation in the objective of Network Utility Maximization (NUM). Significantly, he showed that both centralized and decentralized pricing algorithms were capable to achieve NUM. In [8], Dharwadkar et al. studied the utility functions from the point of their shapes. They categorized the utility functions into three general types: step, linear and concave, and based on these features proposed a heuristic scheduling algorithm that executed dynamic bandwidth allocation and achieved Network Utility Maximization (NUM). Zimmermann and Killat argued in [17] that the utility function stands for the user's preference of bandwidth, which can be modeled as an increasing, strictly concave, and continuously differentiable function, perfectly fitted by the logarithm function. The utility function of HTTP-like traffic class was studied in [7] by Chang et al. They derived a close-form expression for the utility function of HTTP traffic from the behavior of HTTP applications and underlying TCP connections. Harks and Poschwatta [10] proposed scheduling algorithms under the "utility fair" assumption, where bandwidth is allocated such that each user is offered with equalized utility to guarantee fairness. Massoulie and Roberts [15] generalized three objectives for bandwidth allocation in network links: max–min fairness, proportional fairness and minimum potential delay. They developed corresponding scheduling algorithms for each of them respectively.

Although the previous works contributes a lot in building up the basic theoretical framework of the utility-based scheduling and bandwidth allocation, as well as the concerning pricing strategy, at this time, no single work has emphasized on the practical issue of scheduling triple-play services under the background of NGN. Motivated by the desire to bridge such a gap between theory and reality, we work through this issue with the well-known NUM objective. By classifying NGN traffic into five categories according to their diversified utility functions, we explicitly formulate this issue into a nonlinear programming problem with both inequality and equality constraints. After some safe approximations, we further translate it into a nonlinear programming problem with only equality constraints, which can be solved accurately by adopting the well-known Lagrange Multiplier method.

We discuss the solution using our theoretical method in a simplified scenario where there only exist IPTV users and TCP elastic users. A new feature for the NUM-based scheduling is discovered that the IPTV users will face a turning point in bandwidth allocation, before which they actually gain no bandwidth at all.

Simulations on the original scheduling problem under two network scenarios are carried out using nonlinear programming software LINGO. Some bandwidth allocation results under NUM-based scheduling are observed: (1) In both network scenarios, the utilities of VoIP users and other low-throughput real-time UDP users are well guaranteed regardless of the network provisioning conditions, since they are the most cost-effective traffic for bandwidth allocation; (2) IPTV users give up all the bandwidth at first when network provisioning is below a turning point and after that step directly to nearly half of its maximal bandwidth requirement; (3) TCP elastic and interactive users are provisioned nearly proportionally except around IPTV user's turning point where the bandwidth allocation is rather unstable; (4) NUM-based scheduling in general achieves at least 25% utility gain over the strict-priority based scheduling.

We also discuss two measures to alleviate the impact of IPTV's turning point: one is to increase the penetration rate of IPTV service and the other is to elevate IPTV user's maximal utility. Through simulations, we find that the latter method is more practical under current Internet environment.

Compared with the strict-priority scheduling deployed extensively in industry, our results demonstrate that while offering highest strict priority for VoIP users is indeed the best choice, assigning IPTV users the second-highest strict priority actually does not accord well with the objective of NUM. In highly congested networks, the utility gain in allocating bandwidth to IPTV users is rather limited since IPTV user has a considerably high bandwidth threshold to be well provisioned.

The rest of this paper is organized as below. In Section 2, we present our NGN traffic classifications and formulate the utility function for each class. In Section 3, we solve the equivalent nonlinear programming problems. In Section 4, we compute numerical results by LINGO. Section 5 discusses some limitations of this paper and points out future directions. Finally, Section 6 concludes the paper.

2. NGN user classifications and their utility functions

Due to the remarkable distinction in QoS requirements among NGN users, it is important for designers to understand their classifications and treat different traffic classes disparately to achieve global welfare maximization. In this paper, we partition NGN users into five categories according to their explicit QoS requirements: (1) voice over Internet Protocol (VoIP) users previously using traditional PSTN; (2) emerging Internet Protocol Television (IPTV) users; (3) traditional TCP elastic users, including those relying FTP or P2P to download files; (4) web users and other TCP interactive users; (5) other streaming and gaming users generating UDP traffic.

To measure user satisfaction degree, we introduce the well-known concept of user utility, which is first invented

in economic scopes and then borrowed to networking studies ever since [16]. User utility can be quantified as the user's willingness to pay, i.e., the maximal amount of money he would like to give up in exchange of the service. Here, we assume that the user utility of each traffic class is determined only by the allocated equivalent bandwidth [9], which forms the intensively used term of utility function. Although the impacting factors on user utility are actually multi-fold, e.g., delay, delay jitter and loss rate, all the other metrics can be calculated from the allocated bandwidth if the scheduling algorithm is pre-defined.

In the following of this section, we illustrate the utility function of each traffic class in NGN, according to detailed traffic investigations and previous studies in this area. These utility functions build up a fundamental basis for our further optimization solution.

2.1. VoIP user

Since VoIP application is extremely sensitive to packet delay and loss caused by bandwidth insufficiency, its utility function falls into the category of hard real-time class [6,8,16], with a minimal bandwidth requirement of $B_{\min 1}$. When the allocated bandwidth is less than $B_{\min 1}$, user utility will drop to zero. Here, we set $B_{\min 1}$ to 64 Kbps, which is the standard voice encoding rate in most wired telephone communications.

To quantify user utility, we denote the maximal utility of each VoIP user as V_1. Then the utility function at user bandwidth of b_1 can be determined by the stair function

$$u_1(b_1) = V_1 \cdot \frac{\mathrm{sgn}(b_1 - B_{\min 1}) + 1}{2} \quad (b_1 \geqslant 0) \tag{1}$$

As shown in Fig. 1, V_1 actually stands for the importance of VoIP user's traffic compared with the other traffic classes. We use similar denotations to identify this preference in defining other utility functions.

2.2. IPTV user

As a real-time application, IPTV user's utility function is similar to VoIP user's. There exists a minimal encoding rate for IPTV transmission, denoted as $B_{\min 2}$. When bandwidth provisioning for IPTV user is below $B_{\min 2}$, its utility is closed to zero. However, the difference from hard real-time application is, when bandwidth provision increases beyond $B_{\min 2}$, its utility does not increase directly to V_2 (the maximal utility of each IPTV user). Instead, there is a slope area, from $B_{\min 2}$ to $B_{\max 2}$, when user utility adds up smoothly to V_2. The shape of IPTV traffic's utility function is depicted in Fig. 2. We model it with the "Logistic Model" [5] reflecting a typical S-type curve. Correspondingly, IPTV user utility function is quantified by

$$u_2(b_2) = V_2 \cdot \frac{1}{1 + (1/\varepsilon - 1)e^{-r_2 b_2}} \quad (0 \leqslant b_2 \leqslant B_{\max 2}) \tag{2}$$

$$r_2 = 2\ln(1/\varepsilon - 1)/B_{\max 2} \tag{3}$$

For detailed derivation of (2) and (3), please refer to Appendix A. Here, b_2 denotes the actual bandwidth allocated to each IPTV application, ε denotes the tiny IPTV user utility when allocated bandwidth is $B_{\min 2}$. It also determines to what extent IPTV user's utility function is close to the hard real-time one. As ε is set smaller, IPTV traffic is closer to the hard real-time traffic.

According to the latest IPTV encoding format, such as MPEG-2 and MPEG-4/H.264, the bandwidth requirement is normally multi-megabits per second, e.g., in MPEG-2, regulated to 5–15 Mbps, and in MPEG-4/H.264, larger viable range is allowed, from multi-hundreds Kbps to multi-hundreds Mbps. Here, we choose a typical setting of $B_{\min 2} = 100$ Kbps, $B_{\max 2} = 10$ Mbps and $\varepsilon = 0.01$.

2.3. TCP elastic user

TCP elastic user implements traditional TCP protocol with built-in congestion control mechanisms. The utility function of these users features an increasing, strictly concave and continuously differentiable curve [17], which has decreasing marginal increment as bandwidth increases [16]. The logarithm function is often used to quantify its utility function [12–14], as shown in Fig. 3. Denote $B_{\max 3}$ as the maximal bandwidth requirement of each TCP elastic traffic user, this utility function can be formularized as

$$u_3(b_3) = V_3 \cdot \frac{\ln(b_3 + 1)}{\ln(B_{\max 3} + 1)} \quad (0 \leqslant b_3 \leqslant B_{\max 3} \quad b_3, B_{\max 3} : Mbps) \tag{4}$$

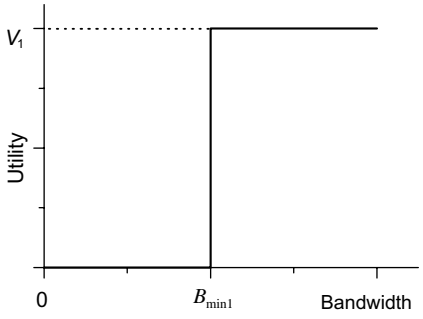

Fig. 1. Utility function of VoIP user.

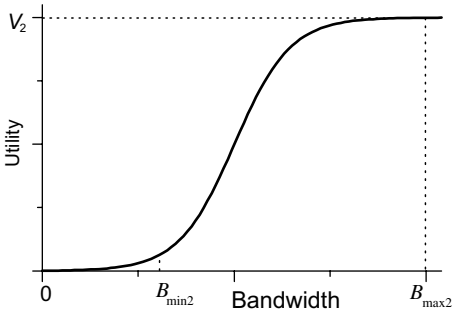

Fig. 2. Utility function of IPTV user.

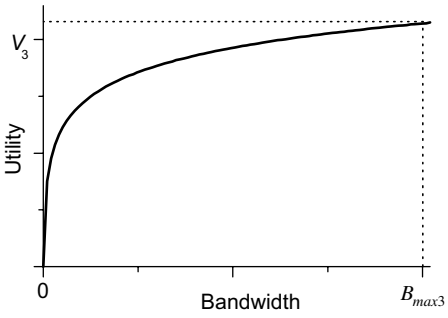

Fig. 3. Utility function of TCP elastic user.

where b_3 denotes the actual bandwidth allocated to each user. We set $B_{\max 3}$ to be a typical value of 10 Mbps in our following analysis.

2.4. TCP interactive user

TCP interactive user mainly includes web users and telnet users who concern packet delays. When web users are surfing in Internet, he will be impatient for waiting long time before retrieving information. For the telnet users, it is also important that the TCP Round Trip Time (RTT) is sufficiently small so that he could feel "connected". In worst case when RTT is larger than some threshold, the telnet session will even be interrupted. Subsequently, the utility function of TCP interactive user, different from TCP elastic one, has a minimum tolerable bandwidth $B_{\min 4}$, below which the utility drops directly to zero.

We depict TCP interactive user's utility function in Fig. 4 and derive its expression as (5).

$$u_4(b_4) = V_4 \cdot \frac{\ln(b_4/B_{\min 4})}{\ln(B_{\max 4}/B_{\min 4})} \frac{\operatorname{sgn}(b_4 - B_{\min 4}) + 1}{2} \quad (0 \leqslant b_4 \leqslant B_{\max 4}) \tag{5}$$

Here b_4 denotes the actual bandwidth allocated to each TCP interactive application. $B_{\max 4}$ denotes the maximal bandwidth requirement, beyond which TCP interactive user will gain little utility increment. We take a typical setting of $B_{\min 4} = 64$ Kbps and $B_{\max 4} = 10$ Mbps.

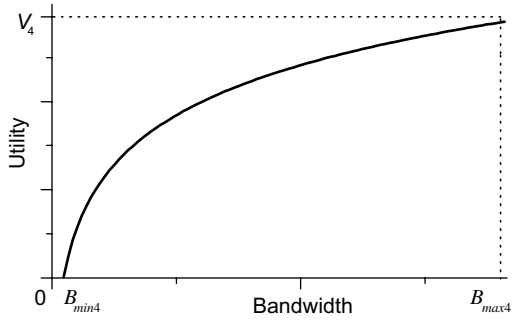

Fig. 4. Utility function of TCP interactive user.

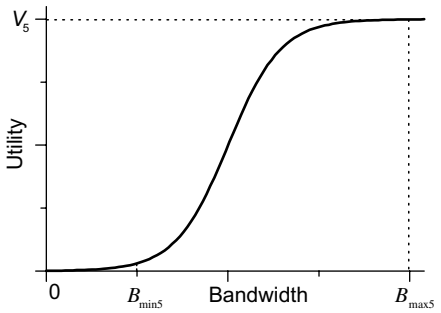

Fig. 5. Utility function of UDP user.

2.5. UDP user

Domain Name System (DNS) users, streaming media users as well as on-line gaming users all belong to this class. Featuring its delay-sensitive characteristic, UDP user's utility function resembles that of IPTV user, with a maximal bandwidth requirement $B_{\max 5}$, beyond which reaches the maximal utility, and a minimal bandwidth requirement $B_{\min 5}$, below which gains no utility at all. However, UDP user's bandwidth demand is far less than IPTV user. For streaming applications, the required bandwidth is from multi-hundred Kbps to multi-Mbps. For on-line gaming applications, many reports [2,4] indicate that a bandwidth of tens of Kbps is capable enough to guarantee a smooth running of most games. Knowing that every application type in this class has a utility function similar to IPTV user, as given in (2), we can derive the UDP user's average utility function. Assume that there are n UDP user types occupying the internal traffic proportions of $p_{51},\ldots,p_{5i},\ldots,p_{5n} (1 \leqslant i \leqslant n)$, which satisfy $\sum_{i=1}^{n} p_{5i} = 1$. We also denote their maximal bandwidth requirements as $B_{\max 51},\ldots,B_{\max 5i},\ldots,B_{\max 5n} (1 \leqslant i \leqslant n)$, then the expression of UDP user's average utility function is calculated by

$$u_5(b_5) = V_5 \cdot \frac{1}{1 + (1/\varepsilon - 1)\mathrm{e}^{-r_5 b_5}} \quad (0 \leqslant b_5 \leqslant B_{\max 5}) \tag{6}$$

$$r_5 = 2\ln(1/\varepsilon - 1)/B_{\max 5} \quad (1 \leqslant i \leqslant n) \tag{7}$$

$$B_{\max 5} = \sum_{i=1}^{n} p_{5i} B_{\max 5i} \tag{8}$$

$$r_5 = 1/\sum_{i=1}^{n} (p_{5i}/r_{5i}) \tag{9}$$

For detailed derivation of (6)–(9), please refer to Appendix B. Here, b_5 denotes the average bandwidth allocated to each UDP application. $B_{\max 5}$ denotes the maximal bandwidth required by each UDP users on average. We find that UDP user's utility function still holds the same format with (2), as depicted in Fig. 5. We set $B_{\max 5}$ to a typical value of 500 Kbps and ε to 0.01 as IPTV utility function.

3. Network utility maximization

Based on NGN user's utility functions, we can solve the congestion-phased bandwidth allocation problem while

conforming to NUM. Consider a single link L with bandwidth C, denote the number of NGN users utilizing L as N and the proportion of five user types as p_1, p_2, p_3, p_4, p_5, corresponding to VoIP, IPTV, TCP elastic, TCP Interactive and UDP users respectively, satisfying $p_1 + p_2 + p_3 + p_4 + p_5 = 1$. We assume that the users of same type will be provided with equal utility according to the utility-fair assumption. Hence, the total utility of user type i can be written as $U_i = Np_iu_i(b_i)$, where b_i is the average bandwidth provided for each user of type i. We also assume that the users belonging to the same type have the identical utility function, then the utility-fair bandwidth allocation inside each type of user group also leads to the bandwidth-fair allocation. That is, the bandwidth allocated for each user of the same type is identical.

Therefore, the aggregate bandwidth occupied by all the user types can be summed as $N\sum_{i=1}^{5} p_i b_i$, and the total utility gain on this link is formalized by

$$U = N\sum_{i=1}^{5} p_i u_i(b_i) \tag{10}$$

The bandwidth allocation issue can therefore be equalized to the nonlinear programming problem with both equality constraints and inequality constraints.

Nonlinear programming problem 1 (with inequality and equality conditions)

$$\begin{aligned}
\text{Minimize} \quad & -U = -N\sum_{i=1}^{5} p_i u_i(b_i) \\
\text{Subject to} \quad & N\sum_{i=1}^{5} p_i b_i = C \\
& 0 \leqslant b_1 \leqslant b_{\max 1} \\
& 0 \leqslant b_2 \leqslant b_{\max 2} \\
& 0 \leqslant b_3 \leqslant b_{\max 3} \\
& 0 \leqslant b_4 \leqslant b_{\max 4} \\
& 0 \leqslant b_5 \leqslant b_{\max 5}
\end{aligned}$$

However, it is difficult to solve this nonlinear programming problem with inequality conditions. Below, we further approximate it into a solvable nonlinear problem with only equality conditions and then discuss the possible solutions.

First, we extend the bandwidth constraint on each utility function from $0 \leqslant b_i \leqslant b_{\max i}$ to $-\infty < b_i < +\infty$, and redefine these utility functions by

$$u_i^*(b_i) = \begin{cases} -\infty & b_i < 0 \\ u_i(b_i) & 0 \leqslant b_i \leqslant b_{\max i} \\ V_i & b_i > b_{\max i} \end{cases} \tag{11}$$

If we consider the situation when link L is congested, i.e., $N\sum_{i=1}^{5} p_i b_{\max i} > C$, there will be no solution to the new nonlinear programming problem (as below) having $b_i < 0$ or $b_i > b_{\max i}$, since (1) in former case, the total utility gain will drop below zero; (2) in latter case, there exists better solution by allocating the bandwidth beyond $b_{\max i}$ to other non-satiated user types. Therefore, the inequality conditions in Problem 1 become inactive in our case, so we can convert it to the problem below.

Nonlinear programming problem 2 (equality conditions)

$$\begin{aligned}
\text{Minimize} \quad & -U = -N\sum_{i=1}^{5} p_i u_i^*(b_i) \\
\text{Subject to} \quad & N\sum_{i=1}^{5} p_i b_i = C
\end{aligned}$$

This problem can be solved by introducing the famous Lagrange multiplier λ^*, i.e., the solution $b^* = (b_1, b_2, b_3, b_4, b_5)$ must satisfy

$$N\sum_{i=1}^{5} p_i b_i = C \tag{12}$$

$$\nabla - U(b^*) = (Np_1, Np_2, Np_3, Np_4, Np_5)^{\mathrm{T}}\lambda^* \tag{13}$$

Eq. (13) is equivalent to

$$u_1^{*\prime}(b_1) = u_2^{*\prime}(b_2) = u_3^{*\prime}(b_3) = u_4^{*\prime}(b_4) = u_5^{*\prime}(b_5) = -\lambda^* \tag{14}$$

However, using Lagrange multiplier method requires the objective function to have continuous first-order derivatives, we achieve that by adding buffering curves to the gradient of $U(b^*)$ (composed by $u_i'(b_i)$, $i = 1, 2, 3, 4, 5$) to gain continuity. In Figs. 6–10, we depict the modified functions of $u_i^{*\prime}(b_i)$, $i = 1, 2, 3, 4, 5$. The buffering curves are added around 0^- to connect the utility function from $-\infty$ to 0.

Below, we derive the solution to the programming problem 2 in an illustrative way. As in Figs. 6–10, by setting a fixed one-dimension Lagrange multiplier λ^*, the bandwidth allocation for users of each type can be visualized. Importantly, for each user type, there are multiple choices, e.g., in Fig. 7 for IPTV user's utility function, three possible allocation results (b_2', b_2'' and b_2''') hit the multiplier λ^*. To obtain the optimized solution, the aggregated bandwidth allocation of all users should be exactly the capacity of the underlying physical link following (12). Quantitatively, this requirement can be formalized by

$$\sum_{i=1}^{5} p_i (u_i^{*\prime})^{-1}(-\lambda^*) = C/N \tag{15}$$

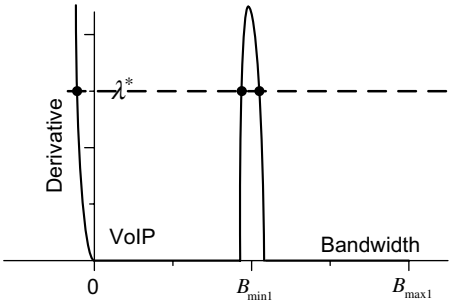

Fig. 6. Derivative of VoIP user's utility function.

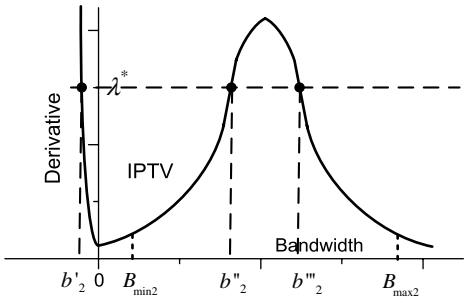

Fig. 7. Derivative of IPTV user's utility function.

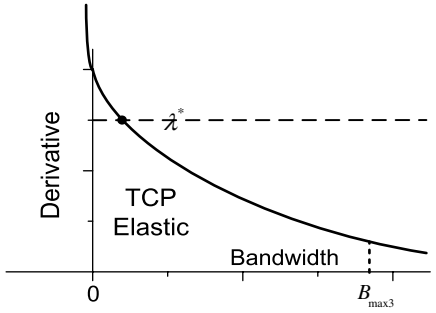

Fig. 8. Derivative of TCP elastic user's utility function.

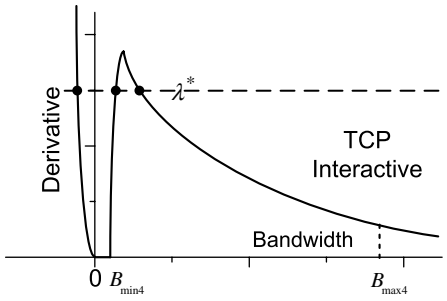

Fig. 9. Derivative of TCP interactive user's utility function.

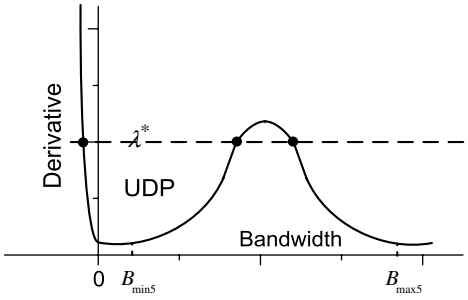

Fig. 10. Derivative of UDP user's utility function.

Since most of the utility functions, except that of TCP elastic user, are not injections, they have multiple inverse functions so that the aggregated inverse function also has several function curves and the solutions to (15) are not exclusive. According to Lagrange multiplier theory, these solutions are all local minimization points, and not necessary to be global minimization point. Hence, we should find out the global minimization point among all the possible solutions by comparing their aggregated user utilities. The one with largest utility is the desired exclusive solution.

Using the method described above, the optimized bandwidth allocation results can be calculated accurately. We further detail our method in solving this problem under a simplified scenario. We assume:

(1) the utility among different user types are fair, i.e., the maximal utility in each user type are of equal value, corresponding to $V_1 = V_2 = V_3 = V_4 = V_5 = 1$; (2) as VoIP user's bandwidth requirement is relatively small, say 64 kbps per user, we fully satisfy its demand before allocating bandwidth among other user types. Distracting bandwidth from VoIP user to other type of users will not increase the aggregated utility, so this simplification does not distort the global optimization solution; (3) since UDP user's utility function is of similar shape with IPTV user, we do not include UDP user in our scenario, the complicated mixed situation is examined through our simulations. For the same reason, we only include TCP elastic user in our analytic scenario and exclude the TCP interactive one.

Then there only remain two types of user in the final optimization phase: the IPTV user and the TCP elastic user. Following the detailed parameters of their utility functions in Section 2, we illustrate the function curve of (15) in Fig. 11, where we set $p_2 = p_3 = 0.5$ and $p_1 = p_4 = p_5 = 0$. Assuming the case that $C/N = 5$ Mbps, there exists three optimization solutions, as shown by the dots in the figure. Using the first solution, as in Fig. 11b, the aggregated user utility is proportional to the size of shadowed area by a multiplier of N. In Fig. 11c which shows the second solution, the total user utility is proportional to the left-shadowed area size minus the right-shadowed area size. And in Fig. 11d which details the third solution, the summed user utility is also proportional to the left-shadowed area size minus the right-shadowed area size.

Since both the first and the third solutions are apparently better than the second one, the final optimization trade-off only happens between them. We then focus on this trade-off. As shown in Fig. 12a, when the bandwidth provisioning for each user is relatively high and the size of left-shadowed area is larger than that of the right-shadowed area, the third solution is the best. In this case, we also illustrate the bandwidth allocation among the two user types where $p_3b_3{:}p_2b_2 = b_3{:}b_2$. In Fig. 12b, when bandwidth provisioning drops to a *turning point* where the sizes of the two areas are identical, it is of equal value to select the first and the third solutions, so the IPTV users can still allocate considerable bandwidth. As the bandwidth provisioning further shrinks to the case in Fig. 12c, all the bandwidth is best allocated to TCP elastic users so as to maximize overall user utility.

This discovery of turning point in NUM based bandwidth allocation is important, since after this point all the IPTV users will be served with no bandwidth at all if the

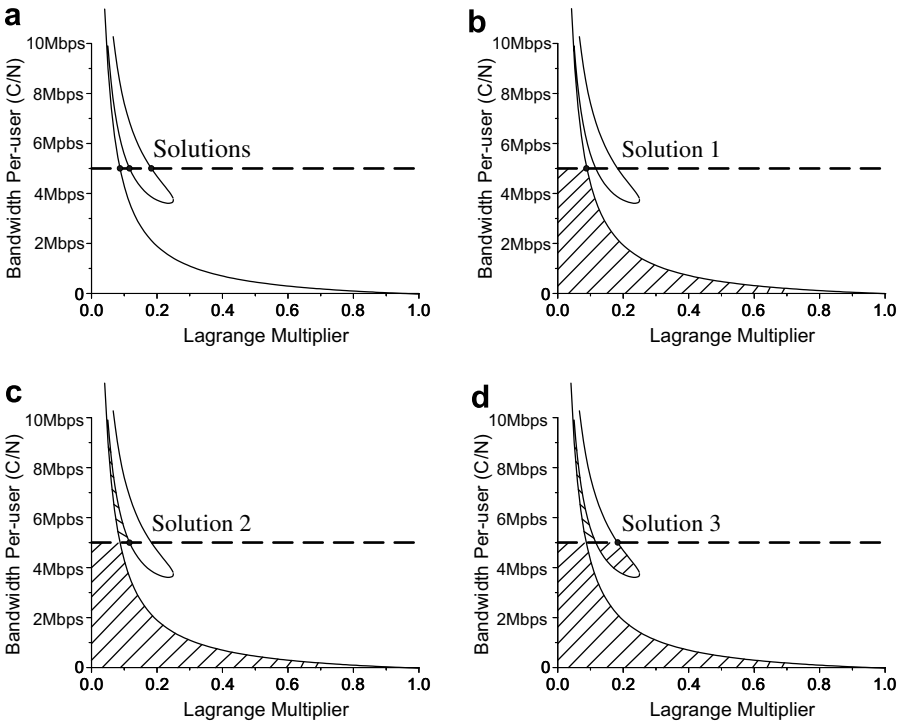

Fig. 11. Solutions for the nonlinear programming problem 2 in the simplified scenario: (a) Three solutions. (b) Solution 1 and its utility. (c) Solution 2 and its utility. (d) Solution 3 and its utility.

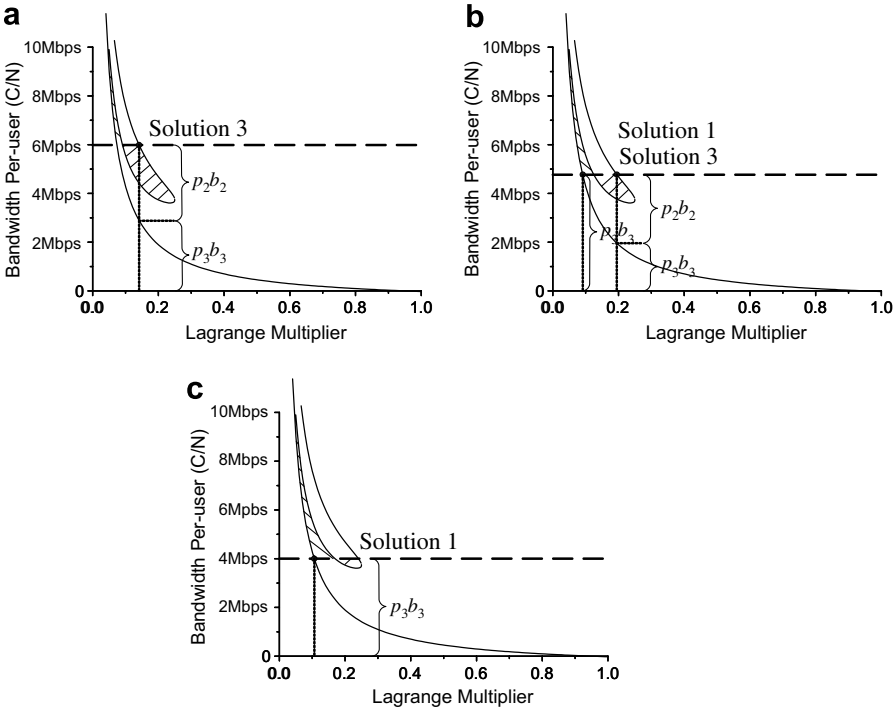

Fig. 12. Impacts of the bandwidth provisioning: (a) When bandwidth provisioning is above IPTV's turning point. (b) When bandwidth provisioning is just at IPTV's turning point. (c) When bandwidth provisioning is below IPTV's turning point.

Internet planner intends to optimize global welfare. Apparently, IPTV users do not at all want this happen. Below, we discuss two methods to alleviate the impact of turning point.

(1) We increase the proportion of IPTV users and find out that the turning point drops correspondingly. In Fig. 13a, we illustrate the turning point when $p_2 = 1/3$ and $p_3 = 2/3$; in Fig. 13b, we draw the case when

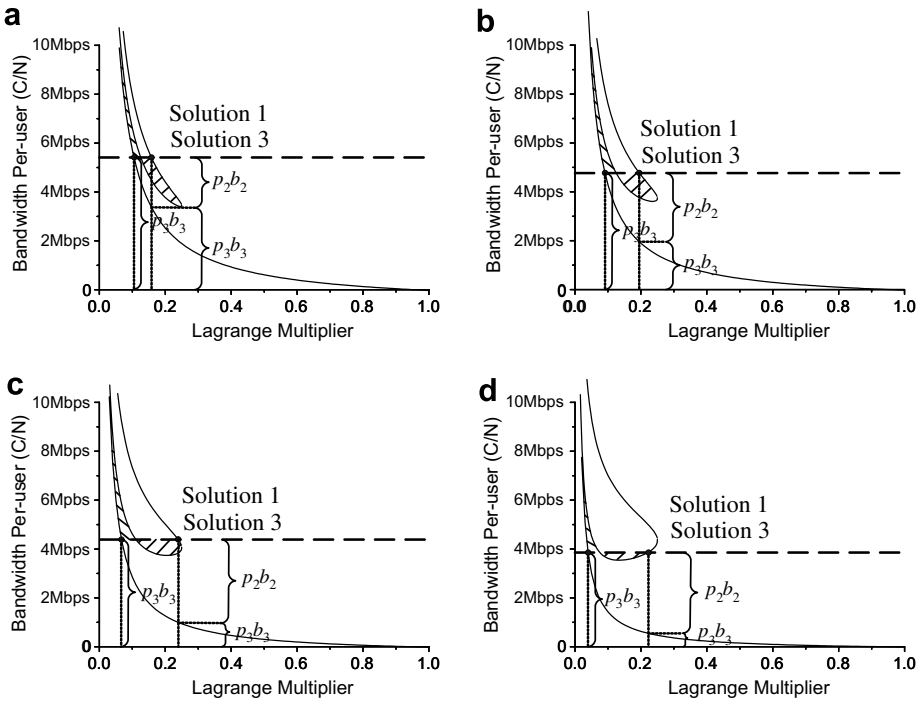

Fig. 13. Turning point positions as IPTV user proportion increases (p_2 denotes IPTV user proportion, p_3 denotes TCP elastic user proportion): (a) $p_2 = 1/3$ and $p_3 = 2/3$. (b) $p_2 = p_3 = 0.5$. (c) $p_2 = 2/3$ and $p_3 = 1/3$. (d) $p_2 = 5/6$ and $p_3 = 1/6$.

$p_2 = p_3 = 0.5$; in Fig. 13c, we depict the curves when $p_2 = 2/3$ and $p_3 = 1/3$; finally in Fig. 13d, the situation with $p_2 = 5/6$ and $p_3 = 1/6$ is illustrated. As the proportion of IPTV user increases, the "hook" in these figures becomes fatter, hence the turning point is lowered, leading to better tolerance for IPTV services to the network congestion.

(2) We increase the maximal utility of IPTV user (V_2) and keep that of TCP elastic user (V_3) unchanged. The proportions of them are still set to $p_2 = p_3 = 0.5$. The illustrations of turning point under $V_2 = 1, 2, 4, 8$ are shown in Fig. 14a–d, respectively. It is observed that the bottom of the "hook" becomes wider as V_2 increases, therefore the turning point drops correspondingly.

Summarily, if the IPTV users want to acquire bandwidth without breaking the NUM rules, there are two measures to take or situations to wait: the first one is to

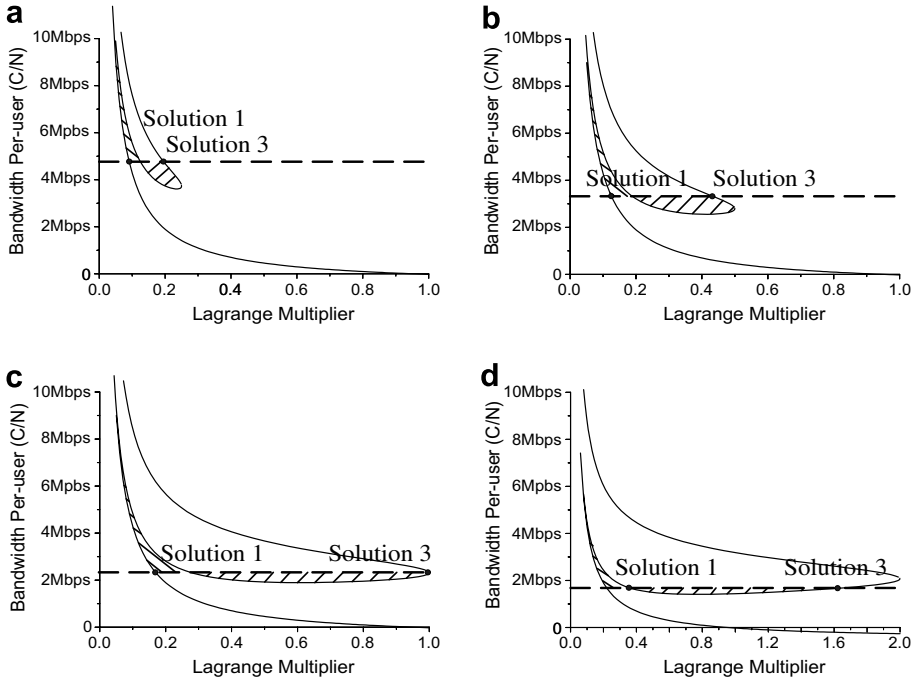

Fig. 14. Turning point positions as IPTV user's maximal utility (V_2) increases: (a) $V_2 = 1$. (b) $V_2 = 2$. (c) $V_2 = 4$. (d) $V_2 = 8$.

increase IPTV user's proportion, in other words, wait for a higher penetration of IPTV services; the other one is to boost the utility of IPTV services, i.e., pay more from the viewpoint of both users and service providers.

In the following simulation section, we will further examine the affection of above factors to the NUM based bandwidth allocation. It is clearly demonstrated that the last method, to pay more, is a more effective way to advance the turning point of IPTV users.

4. Simulations and numeric results

In this section, we present the numeric results of bandwidth allocation based on NUM by the nonlinear programming software LINGO. We also simulate the allocation result under the strict-priority scheduling for comparison. In the latter scheduling, VoIP traffic is given the highest priority, followed by IPTV traffic, TCP interactive traffic, UDP traffic and TCP elastic traffic with decreased priorities.

We investigate bandwidth allocation results under two network scenarios. The first one is the current Internet, where HTTP and TCP elastic traffic still dominate in traffic volume, namely data-dominated network; the other is the prospective NGN, where the emerging services, especially the IPTV traffic will be responsible for most of traffic, namely the IPTV-dominated network. In data-dominated network, the proportions of the five user types, i.e., VoIP, IPTV, TCP elastic, TCP Interactive and UDP are set to {10%, 10%, 10%, 50%, 20%}; while in IPTV-dominated network, we set them to {10%, 50%, 10%, 20%, 10%}.

We also investigate the turning point position for IPTV users under two cases: (1) as IPTV user's proportion (penetration) grows; (2) as IPTV user's maximal utility increases. The simulation results are of some difference with the analytic one in Section 3 since it emulates the entire picture of our bandwidth allocation problem, not only the simplified one in Section 3.

In all the simulations, we set the capacity of the congestion-phased network link to be $C = 10$ Gbps, and tune the user number on this link to gain the bandwidth allocation results under different bandwidth-per-user parameters.

4.1. Data-dominated network

Fig. 15a shows the bandwidth allocation for user of each type in the data-dominated network when the average per-user bandwidth provisioning increases, i.e., network congestion degree alleviates. Fig. 15b gives the results using strict priority scheduling. Comparing the two figures, the most significant differences of NUM-based scheduling from strict priority scheduling lies in that: (1) it is more like the proportional bandwidth allocation algorithm except that some user types, e.g., VoIP and UDP, are fully satisfied from the beginning; (2) there exists a turning point for IPTV users that before this point they are offered with no bandwidth at all and after this point they are immediately given more than half of their desired bandwidth. The scheduling near the turning point of IPTV users is rather unstable that may degrade user satisfaction.

We have calculated the average user utility under these two scheduling approaches and drawn the comparison in Fig. 17a. It shows that the strict-priority based scheduling suffers from a utility loss of at least 20% in most cases. (corresponding to the utility gain of 25% for NUM-based scheduling.)

4.2. IPTV-dominated network

Fig. 16a shows the bandwidth allocation for user of each type in the IPTV-dominated network when the average per-user bandwidth provisioning is increased. Besides it, Fig. 16b depicts the allocation results using strict-priority scheduling in the same scenario. We find that in IPTV-dominated network, there still exists turning point for IPTV users before which they receive no bandwidth at all. Something difference under this situation is that because of the increase of IPTV user proportion, after

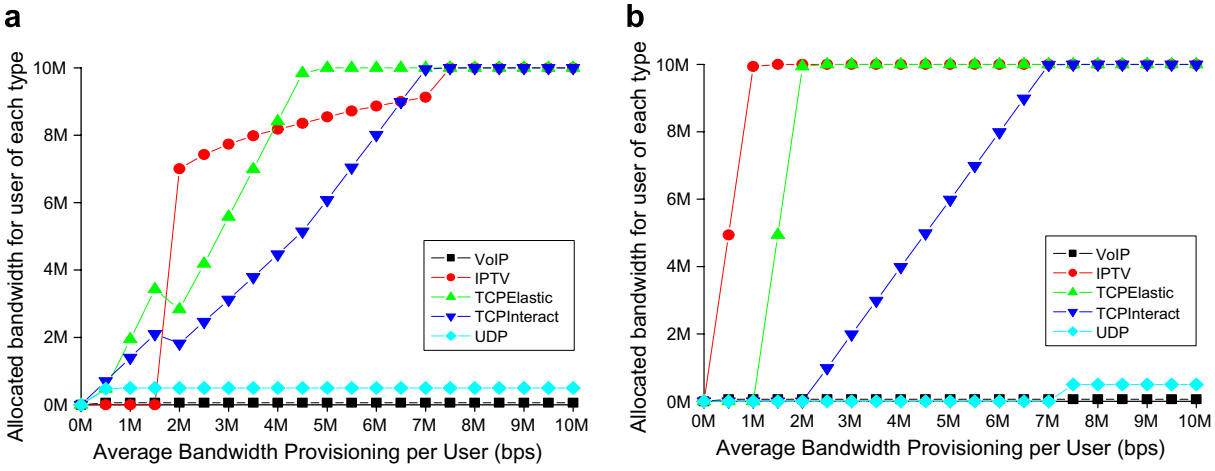

Fig. 15. Bandwidth allocation results in data-dominated network: (a) Under NUM objective. (b) Using strict-priority scheduling.

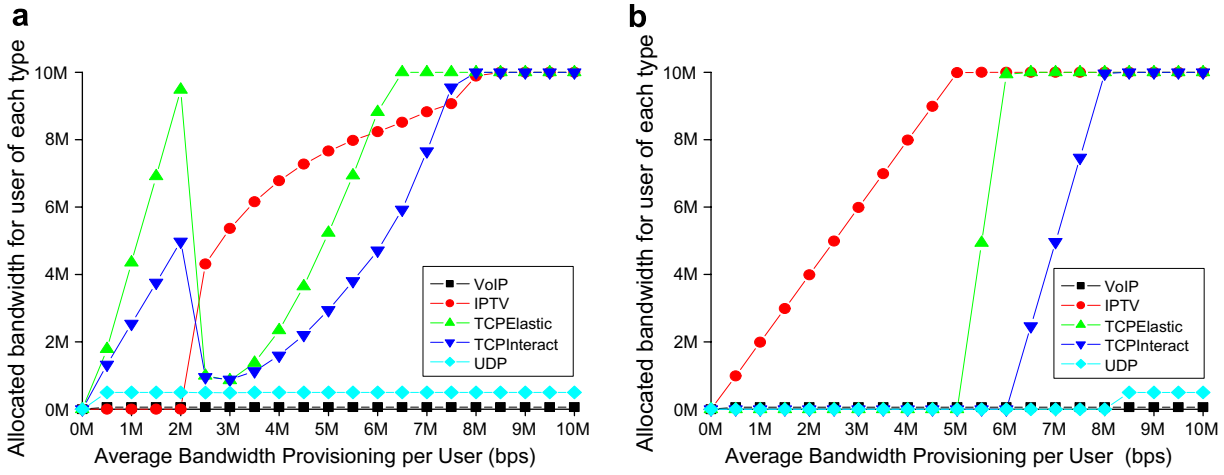

Fig. 16. Bandwidth allocation results in IPTV-dominated network: (a) Under NUM objective. (b) Using strict-priority scheduling.

the turning point, the other TCP users (including TCP elastic and interactive users) suffer from a huge drop in bandwidth allocations. This adds up to the scheduling instability around the IPTV's turning point.

Fig. 17b presents the utility loss of strict-priority scheduling. Under most congested situations (average bandwidth per-user below 6 Mbps), this utility loss is more than 25% (corresponding to the utility gain of 33% for NUM-based scheduling).

4.3. Increase IPTV penetration

Fig. 18 depicts the vicissitude of IPTV user's turning point. As IPTV penetration grows, we find the turning point at first delayed (enlarging the zero-bandwidth area) and then after IPTV users occupy more than 60% market, the turning point starts to decrease, shrinking the zero-bandwidth area for IPTV users.

4.4. Scale IPTV utility function

Another method to alleviate the impact of IPTV turning point is to increase its maximal utility. Fig. 18b depicts the effects of such measure. As IPTV user's maximal utility increases, the turning point advances significantly to nearly 0.5 Mbps, greatly limiting the zero-bandwidth area for IPTV users.

5. Discussions

5.1. Implementation issues

The NUM-based scheduling proposed in this paper can be implemented in both the core and the edge routers serving triple-play services. This specially designed traffic scheduler can be equipped in both the input and the output port of the router to function as the traffic-class sub-scheduler in its multi-tiered packet scheduler.

However, to make such a scheduler feasible, one may need to solve the nonlinear programming problem online, which is quite costly according to the analysis in Section 3. One solution is to bypass the computation to offline and apply simplified scheduling in the online scheduler. We provide two methods to achieve that under the scenario considered throughout the paper.

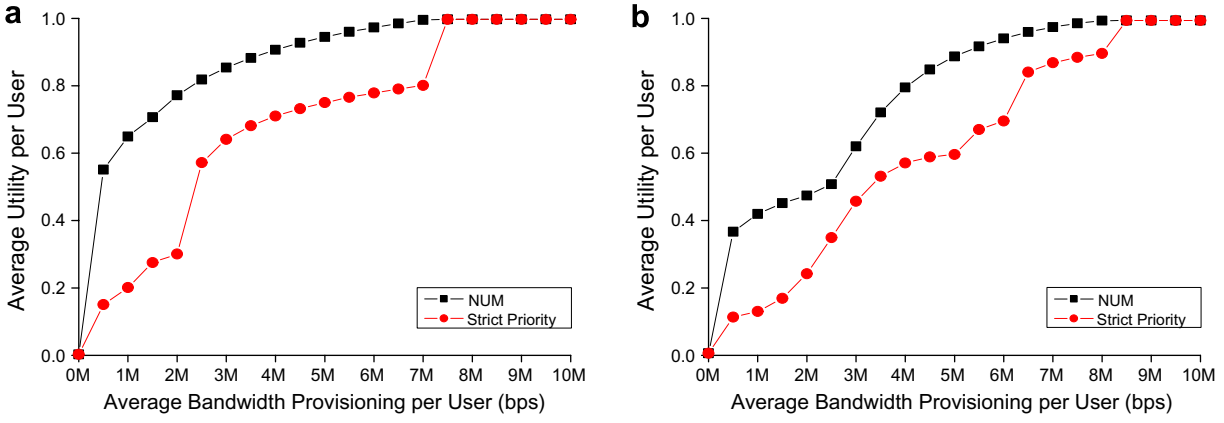

Fig. 17. Utility comparison under NUM-based scheduling and strict-priority based scheduling: (a) In data-dominated network. (b) In IPTV-dominated network.

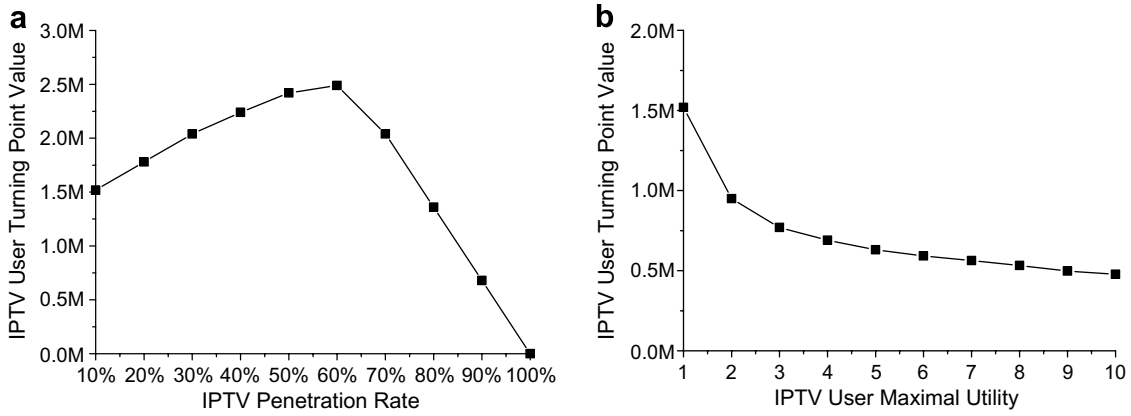

Fig. 18. Turning point position of IPTV users in data-dominated network: (a) As IPTV user proportion increases. (b) As IPTV user's utility increases.

First, according to the results derived in this paper, each traffic class (user type) will have an intrinsic turning point over the bandwidth provisioning condition, before which it will not gain bandwidth if conforming to the NUM objective. These turning points can be calculated offline using the analytical method presented in the paper. We also observe from the simulation results that before and after the IPTV traffic class's turning point, the bandwidth allocation among all the service type is approximately proportional, then we can design two proportional bandwidth schedulers and choose one to use according to the current bandwidth provisioning condition. This method is simple, although not optimal, and is only shown to work for the scheduling of triple-play services.

The second method is to compute the detailed NUM-based bandwidth allocation under each provisioning condition and each traffic pattern (i.e., the proportion of users belonging to each traffic class), and store the results into tables to lookup online. In this way, the timing complexity will be low enough for online scheduling, but the space complexity may be considerably high. One can use compression to reduce the result table size. The details will be out of the scope of this paper.

5.2. Intuitions, limitations and future works

Our results show that NUM-based scheduling differs a lot with strict-priority based scheduling intensively used in current network nodes, with at least 25% utility gain in most cases. The most important finding of our study lies in that to stick to NUM rules, IPTV users must give up all their bandwidth when the average bandwidth provisioning is below a turning point, mostly because their utility functions are not concave before that. Moreover, due to the burst of IPTV bandwidth provisioning, the other user types, such as TCP elastic users and TCP interactive users, have to suffer from unstable bandwidth allocations as well, which in a whole, brings about inconsistent network behavior and user perception. To alleviate this issue, we discuss two methods to minish the effect of turning point. The first one is to increase the IPTV user penetration and the other one is to increase the IPTV user's utility. However, the first method requires the IPTV penetration rate to increase to more than 60%, which is not practical under current Internet environments. Another pertinent solution is to elevate the charging of IPTV services so that the IPTV user with smaller utility will quit the market, leaving only the ones with higher utility, hence indirectly lowers the turning point of IPTV users and reduces their zero-allocation area.

Despite of the clearness of this paper to identify the basic principles to achieve NUM under NGN running triple play services, there are still some leave-outs in our study that may affect the results. First, we adopt the simplified model of utility functions determined only by allocated bandwidth. In real Internet, the impacting factors on the user utility may be complicated, some others includes the bandwidth (sending rate) jitter, one-trip delay and RTT. Second, we assume the user utility is only determined by the scheduling on the studied network nodes, while in real worlds, it is co-affected by those along its packet-transmitting path.

Another limitation of this paper is that we only consider the scheduling decision in objective of NUM, while in real Internet, ISP would only like to maximize its own revenue; Internet users are selfish who only care about their own utilities; there is actually no merciful planner who is able to control the entire network to optimize global welfare.

Therefore, our future works will be studying the bandwidth allocation issue given more sophisticated utility functions and considering the gaming strategy between the service providers and NGN users. However, our final goal remains to maximize the global user utility in the entire NGN.

6. Conclusions

In this paper, we study the problem of scheduling and bandwidth allocation for triple-play services in the objective of NUM. Through generalizing the utility functions of five traffic classes inside NGN, we explicitly solve the equivalent nonlinear programming problem and present theoretical method to compute the bandwidth allocation results. Both this method and the nonlinear programming

software are applied to derive numerical results under two network scenarios. Our results indicate several features of such bandwidth allocations: (1) the VoIP and other low-throughput UDP users can always be guaranteed of sufficient bandwidth; (2) as network congestion becomes severer, IPTV user's bandwidth decreasing encounters a quickly lessening turning point, which is quite concerned with IPTV user's proportion in the network and the utility differentiation among user types; (3) TCP elastic and Interactive users are provisioned nearly proportionally except the instabilities around IPTV's turning point.

Appendix A. Derivations of IPTV user's utility function

We use Logistic model to represent IPTV user's utility function, which can be written as

$$u_2(b_2) = \frac{1}{1 + p\mathrm{e}^{-qb_2}} \quad (p, q > 0) \tag{16}$$

Here b_2 is the actual bandwidth allocated to each IPTV user and u_2 is the corresponding utility. As discussed in the paper, when the bandwidth allocated to IPTV user is $B_{\min 2}$, the utility should be as small as zero, say ε, and when the bandwidth provision is $B_{\max 2}$, the utility should be closed to nearly 1, without loss of generality, say $1-\varepsilon$. Then we have

$$u_2(b_2) = \frac{1}{1 + p\mathrm{e}^{-qB_{\min 2}}} = \varepsilon \tag{17}$$

$$u_2(b_2) = \frac{1}{1 + p\mathrm{e}^{-qB_{\max 2}}} = 1 - \varepsilon \tag{18}$$

Solving (17) and (18), we obtain

$$p = (1/\varepsilon - 1)^{\frac{B_{\max 2} + B_{\min 2}}{B_{\max 2} - B_{\min 2}}}, \quad q = 2\ln(1/\varepsilon - 1)/(B_{\max 2} - B_{\min 2}) \tag{19}$$

Since $B_{\min 2}$ is negligible compared with $B_{\max 2}$, p and q can be further approximated by

$$p = (1/\varepsilon - 1), \quad q = 2\ln(1/\varepsilon - 1)/B_{\max 2} \tag{20}$$

Then IPTV user's utility function can be written as

$$u_2(b_2) = V_2 \cdot \frac{1}{1 + (1/\varepsilon - 1)\mathrm{e}^{-r_2 b_2}} \quad (0 \leqslant b_2 \leqslant B_{\max 2}) \tag{21}$$

where $r_2 = 2\ln(1/\varepsilon - 1)/B_{\max 2}$.

Appendix B. Derivations of UDP user's utility function

Denote the aggregated utility and bandwidth for all the UDP users to be U_5 and B_5, and the number of total UDP users to be N_5. Under assumption that each UDP user is allocated equalized utility of $U_5(B_5)/N_5$, we should have

$$\sum_{i=1}^{n} (N_5 p_{5i}) u_{5i}^{-1}(U_5(B_5)/N_5) = B_5 \tag{22}$$

where u_{5i} is the utility function of the ith application type for UDP users. Then we get

$$U_5(B_5) = N_5 \left(\sum_{i=1}^{n} N_5 p_{5i} u_{5i}^{-1} \right)^{-1} (B_5) \tag{23}$$

By calculating the reverse function for u_{5i} we further obtain

$$U_5(B_5) = N_5 \left(\sum_{i=1}^{n} \left(N_5 p_{5i} \left(-\frac{1}{r_{5i}} \right) \ln \left(\frac{1}{1/\varepsilon - 1} \frac{1 - u_{5i}}{u_{5i}} \right) \right) (u_{5i}) \right)^{-1} (B_5) \tag{24}$$

Since every application for UDP users shares equalized utility, we have $u_{51} = u_{52} = u_{53} = u_{54} = u_{55} = u$, yielding

$$U_5(B_5) = N_5 \left(\sum_{i=1}^{n} \left(-\frac{N_5 p_{5i}}{r_{5i}} \ln \left(\frac{1}{1/\varepsilon - 1} \frac{1 - u}{u} \right) \right) (u) \right)^{-1} (B_5) \tag{25}$$

Calculating their reverse functions, we have

$$U_5(B_5) = \frac{N_5}{1 + (1/\varepsilon - 1)\mathrm{e}^{-\left(1/\sum_{i=1}^{n}(p_{5i}/r_{5i})\right)(B_5/N_5)}} \tag{26}$$

Denoting $r_5 = 1/\sum_{i=1}^{n}(p_{5i}/r_{5i})$ and $b_5 = B_5/N_5$, Eq. (26) can be written as $U_5(B_5) = \frac{N_5}{1+(1/\varepsilon-1)\mathrm{e}^{-r_5 b_5}}$. Further introducing V_5, the expected utility function for UDP users can be written as $u_5(b_5) = V_5 \cdot \frac{1}{1+(1/\varepsilon-1)\mathrm{e}^{-r_5 b_5}}$.

References

[1] AN-2000™ IP-DSLAM for Triple Play, White Paper, UTStarcom Inc., 2004.

[2] How much bandwidth would be used by online gaming, Available from: <http://www.mybroadband.co.za/vb/showthread.php?t=20154>.

[3] Optimizing the Broadband Aggregation Network for Triple Play Services, Strategic White Paper, Alcatel, 2005.

[4] QoS Technology Guide for Gamers, Available from: <http://gamer.ubicom.com/guides/qos_technology_guide.html>.

[5] J. Banks, V. Dragan, et al., Chaos: A Mathematical Introduction, Cambridge University Press, 2003.

[6] Z. Cao, E. Zegura, Utility max–min: an application-oriented bandwidth allocation scheme, in: Proc. IEEE INFOCOM, 1999, pp. 793–801.

[7] C.S. Chang, Z. Liu, A bandwidth sharing theory for a large number of HTTP-like connections, IEEE/ACM Transactions on Networking 12 (5) (2004) 952–962.

[8] P. Dharwadkar, H.J. Siegel et al., A Heuristic for dynamic bandwidth allocation with preemption and degradation for prioritized requests, in: Proc. ICDCS, 2001, pp. 547–556.

[9] R. Guerin, H. Ahmadi, et al., Equivalent capacity and its application to bandwidth allocation in high-speed networks, IEEE Journal on Selected Areas in Communications 9 (7) (1991) 968–981.

[10] T. Harks, T. Poschwatta, Priority pricing in utility fair networks, in: Proc. IEEE ICNP, 2005, pp. 311–320.

[11] F.P. Kelly, Charging and rate control for elastic traffic, European Transactions on Telecommunications 8 (1997) 33–37.

[12] F.P. Kelly, Rate control in communication networks: shadow prices, proportional fairness and stability, Journal of the Operational Research Society 49 (1998) 237–252.

[13] S. Kunniyur, R. Srikant, End-to-End congestion control schemes: utility functions, random losses and ECN marks, IEEE/ACM Transactions on Networking 11 (5) (2003) 689–702.

[14] C.M. Lagoa, H. Che, et al., Adaptive control algorithms for decentralized optimal traffic engineering in the internet, IEEE/ACM Transactions on Networking 12 (3) (2004) 415–428.
[15] L. Massoulie, J. Roberts, Bandwidth sharing: objectives and algorithms, IEEE/ACM Transactions on Networking 10 (3) (2002) 320–328.
[16] S. Shenker, Fundamental design issues for the future Internet, IEEE Journal on Selected Areas in Communications 13 (7) (1995) 1176–1188.
[17] S. Zimmermann, U. Killat, Resource marking and fair rate allocation, in Proc. IEEE ICC, 2002, pp. 1310–1314.