Time-of-day Internet-access management by combining empirical data-based pricing with quota-based priority control

S.-I. Chu and S.-C. Chang

Abstract: An empirical data-based design methodology is proposed for Internet-access management to improve congestion, uneven usage and fairness, especially during peak hours, over a free-of-charge or flat-rate network. The design methodology combines time-of-day pricing (TDP) with quota-based priority control (QPC). Core to the design methodology are the innovations in characterising user demand and quota-allocation behaviour with respect to time and pricing. In-depth analyses of empirical data reveal distinctive behaviour patterns of myopic and prudent quota allocations over time and both patterns indicate high preference for peak-hour access. The user models adopt general utility functions and capture how pricing affects user behaviour as prudent or myopic. Preference parameters of users' utility over time are then estimated by collecting easily measurable user volumes. The TDP design problem is formulated and solved as a Stackelberg game. Tested on the empirical data of a 5000-user network, the TDP design leads to significant improvements in peak-hour usage and fairness, peak shaving and load balancing over pure QPC. The methodology requires only two simple and short-period data collections from an operational network and takes about 1 min of CPU time for TDP calculation. Results demonstrate the effectiveness of our design methodology when applied to Internet-access environments with frequent changes.

1 Introduction

There often exists uneven and unfair usage of Internet access, especially during peak hours, over a network environment where the service charge is free or flat rate. For example, consider the dormitory network of National Taiwan University (NTU), where the network management adopts a quota-based priority control (QPC) scheme [1] to control its Internet access traffic. When a user's Internet-access volume exceeds the daily regular service quota, the user's traffic is directed to a lower priority service. Statistics show that during peak hours, the drop rate of regular service is higher than 2.5 Mbps. The average usage of heavy users (8% of the user population) is 12.08 times more than that of all other users. Such observations imply that even under QPC, congestion and unfair usage are still significant. There is a need for a finer management scheme to regulate users' Internet access over time.

Many researchers have studied time-of-day load management for public utilities. Time-of-day pricing (TDP) [2, 3] and peak-load pricing [4] offer an indirect load management mechanism that meets the dual objectives of (i) reducing peak load, and (ii) shifting a portion of the peak load to the base load. Although the idea is simple, the actual price has to be carefully designed to induce user behaviour and to avoid side effects such as peak shifting [5].

Dynamic pricing [6, 7] was proposed for dynamic traffic management over communication networks. MacKie-Mason and Varian [8] introduced an auction-based pricing scheme, called smart market, where the price varies minute-by-minute to reflect the current state of network congestion. This scheme works without knowledge of the explicit user behaviour model but is difficult to implement over the existing network environments. Jin *et al.* [9] proposed simple gradient algorithms to dynamically adjust prices based on congestion. But this approach requires a small amount of steady communications about demand and supply along each route. Paschalidis and Tsitsiklis [10] studied a model for optimal congestion-dependent (dynamic) pricing of network services, and also provided a comparison with static pricing. One of the important conclusions is that TDP was almost as good as congestion-dependent pricing if the static price profile was suitably chosen.

In [11], Shih *et al.* compared static TDP, call-duration pricing and simple congestion pricing over a computer-telephony-service with about 40 students in the dormitories. Users were not charged by real money but limited by a token budget. Experimental results revealed that TDP and call-duration pricing enticed users to talk at a different time or talk shorter. But the users had no change in their behaviour under congestion pricing because they did not perceive when the price would change. Shih *et al.* showed that the TDP is not only simpler to design and implement than dynamic pricing over an existing environment but also more predictable and acceptable to users. However, the setting of price difference between peak and off-peak hours was not clearly mentioned. How to adapt the TDP profile to a frequently changing environment was not put forth either.

doi:10.1049/iet-com:20060521

Paper first received 27th December 2005 and in revised form 8th January 2007

The authors are with the Department of Electrical Engineering, National Taiwan University, Taiwan, Republic of China

E-mail: scchang@cc.ee.ntu.edu.tw

Be it dynamic or static pricing, characterisation of user behaviour is critical to the design. To understand the relationship between user demands and prices, Edell and Varaiya conducted a market and technology trial over ISDN with 70+ participants under Internet Demand Experiment (INDEX) project [12]. Experimental results showed that user demands were very sensitive to price and quality. The user demand model was also constructed as a logarithmic function of prices for various service classes by regression of experimental data. There was no modelling of user budget. In [1], Lin *et al.* studied user demands and responses to the QPC scheme over NTU dormitory networks with 5000+ users. The network administration intended to induce users to shift part of their peak-hour demands to off-peak hours through the quota limitation. However, Lin *et al.* did not explicitly model the behaviours of how users allocate their quota.

This paper proposes an empirical data-based design methodology for TDP to be combined with QPC, called QPC/TDP, for managing the Internet access over time. This pricing approach may induce user behaviour while allowing user's flexibility in allocating quota based on individual demands over time. The novel design methodology includes four steps: (i) pilot experiment and analyses, (ii) empirical user demand model construction, (iii) TDP design using a game theoretic problem formulation and (iv) network performance and user usage prediction by simulation.

Core to the design methodology are empirical data-based user demand modelling and a game theoretic pricing problem formulation. Analyses of the empirical data of QPC indicate two prominent behavioural patterns of user quota allocation: myopic and prudent behaviours. Myopic users use the quota as demanded at the time, regardless of quota limitation, while prudent users allocate the available quota to maximise their daily benefits. Both users significantly contribute to the traffic of peak hours. Whether a user is prudent or myopic is not fixed but depends on the price, the user's demand and the given quota. Analyses show that without differentiating myopic and prudent user behaviours, a pricing policy may lead to either serious congestion in peak-hours or bandwidth under-utilisation in off-peak hours. In modelling user's utility of Internet access, users have different preferences over time. These preference parameters are estimated by novel exploitation of individual user volumes collected from a network with QPC. As the utility function assumes only diminishing returns to scale and continuous differentiability, the empirical data-based user demand models can be quite general.

The pricing problem is formulated as a Stackelberg (leader–follower) game [13], where the network manager is the leader and the users are the followers. The leader maximises the bandwidth utilisation while keeping the average total demand below the available capacity. A follower maximises one's own benefit, either myopic or daily. Solution analyses of the game formulation show that in the case of purely prudent users, the optimal price profile follows the preference trends, while in the case of purely myopic users, this property may not hold because of their short-sightedness.

To assess the effectiveness of QPC/TDP, the design steps above are tested on the empirical data of a 5000-user network. Results show that the peak-hour usage of heavy users and fairness are improved as compared to QPC only. Peak shaving and load balancing are thus achieved. When a new policy is needed, the network manager has to conduct two simple short-period collections. The policy design and evaluation are efficient in computation. These demonstrate the effectiveness of our design methodology for application to a frequently changing network.

The remainder of this paper is organised as follows. Section 2 describes the deficiencies of QPC and the challenges of pricing. In Section 3, the design methodology of QPC/TDP is proposed. Properties of pricing policies are studied in Section 4. Section 5 exploits the empirical data for effectiveness assessment of QPC/TDP. Finally, we conclude this paper in Section 6.

2 Needs for control over time

2.1 Internet access with QPC

To effectively improve unfairness and uneven usage, the NTU network management adopted a control scheme [1] that combines the ideas of quota limitation and priority differentiation. There are two service classes offered: regular and custody. The regular service has a higher priority than the custody service for data transmission. There is a volume quota for each user's regular service to meet majority users' essential demands. The custody service has no quota limit, which allows heavy users to access the Internet at a lower quality. The default class is regular.

Lin *et al.* [1] implemented QPC over the NTU dormitory networks shown in Fig. 1, where there were 5535 users. Only 54 Mbps was allocated to the outbound traffic from the dormitory networks. Interested readers may refer to [1] for more details.

2.2 Needs and challenges for control over time

The deficiencies of QPC lie in its ineffectiveness in peak-hour traffic management. Statistics show that the Internet-access bandwidth is highly utilised and the drop rate is higher than 2.5 Mbps during peak hours of 9 a.m. to 3 a.m. when the daily quota is 1 GB. Analysis of the peak-hour usage shows that the ratio of heavy users' usage to normal and light users' usage is 1308%. Heavy users still occupy most bandwidth for Internet access.

Moreover, empirical data of QPC reveal the behavioural patterns of myopic and prudent users in quota allocation over time. Fig. 2 shows that a myopic user submitted the volume as desired without considering quota limitation. Such myopicity can be observed from the fact that the user used up the quota by 4 p.m. after quota replenishment at 6 a.m. Most of the quota was used in peak hours. Fig. 3 demonstrates a prudent behaviour. The user allocated the available quota to different time slots based on his/her own preference, which mostly concentrated in peak hours. Both myopic and prudent users contributed significantly to the traffic of peak hours. There is an obvious need for a

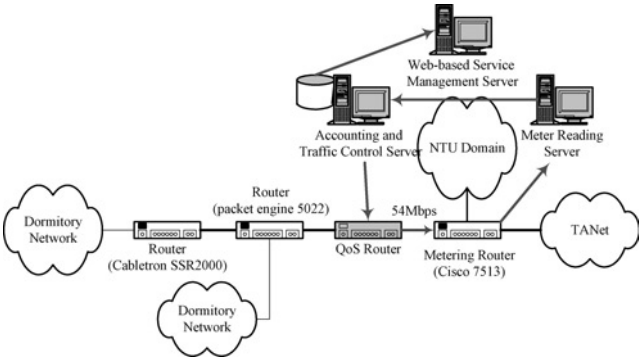


Fig. 1 *Network architecture of QPC*

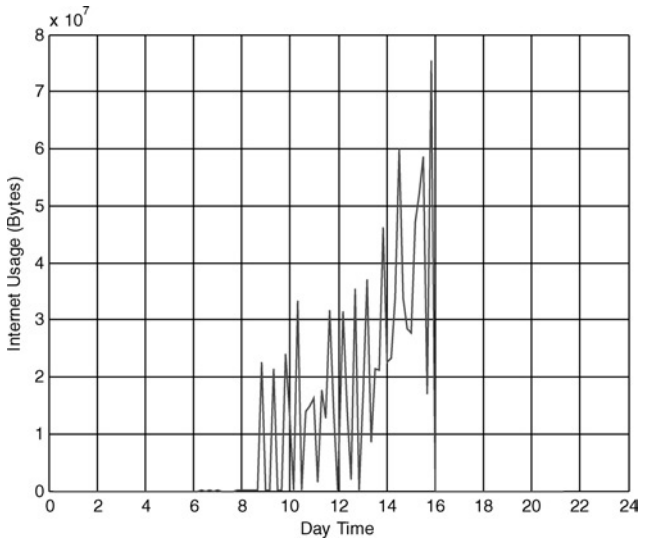


Fig. 2 *Internet access by a myopic user under QPC*

finer time-of-day management scheme to induce users to shift part of their peak-hour demands to off-peak hours.

To improve uneven usage, and peak-hour congestion and fairness, this paper utilises empirical data to model user characteristics and a pricing policy in conjunction with QPC to regulate Internet access over time. This scheme intends to give a user the flexibility in allocating one's quota to demand. From the viewpoint of network managers, they need a methodology to follow up for easy management and require the minimal additions to QPC.

Specific challenges are as follows:

(C1) How to exploit empirical data to model the quota allocation response of myopic and prudent users to pricing over time?
(C2) How to design a pricing policy to be used with QPC to maximise bandwidth utilisation while peak shaving and load balancing effects are achieved?
(C3) How to design a simple and predictable pricing scheme for easy acceptance by users?
(C4) How to exploit the hardware and software of the legacy production network for economic and easy implementation? and
(C5) How to construct a design methodology for application to a frequently changing network environment?

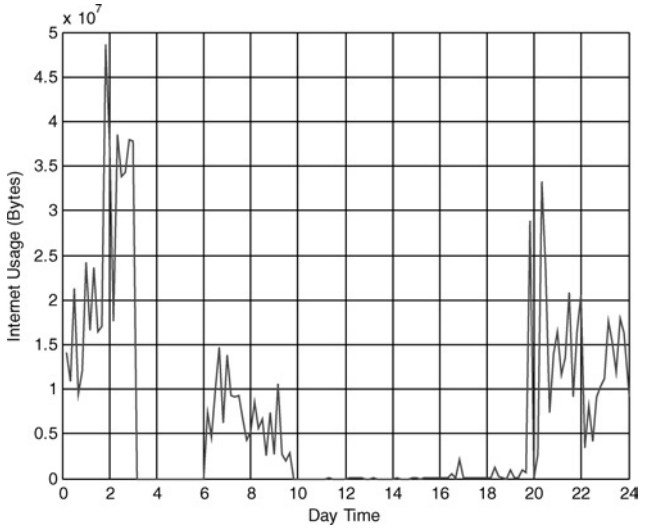


Fig. 3 *Internet access by a prudent user under QPC*

All of the above challenges are addressed in this paper. The beginning of Section 3 addresses (C3) and (C4). Subsequently, the design methodology is proposed for (C5). Section 3.1 constructs quota behaviour models of both prudent and myopic users (C1), and pricing design (C2) is presented in Section 3.2.

3 TDP under QPC

In view of the challenges ((C3) and (C4)) and research conclusions of [10, 11], we adopt a static TDP scheme for Internet access control. Such a simple pricing scheme will allow users to know the price profile in advance so that they can adjust their usage behaviour and will require only an easy addition of a price calculation module to the QPC network management servers. To efficiently design a TDP profile based on empirical network data, a design methodology of QPC/TDP is proposed in Fig. 4. First of all, the network manager has to assess the current network performance (peak-hour drop rate) through monitoring. If the traffic pattern over time changes and the original control policy is no longer effective, the four innovative steps are initiated. (i) The pilot experiment and analysis step includes a baseline experiment and a QPC experiment. First, the baseline experiment has no quota limitation for users and is conducted to characterise the problem of Internet access and user demands over time. The QPC experiment, where all users are allocated one same daily quota, is then conducted to provide data for constructing user demand models. In both experiments, the network manager measures and collects the submitted volumes of individual users at each time slot. (ii) The empirical user demand model construction step models the quota allocation behaviours of myopic and prudent users based on empirical data from the pilot experiments. The model captures the feature that given a quota, whether a user's quota allocation behaviour is myopic or prudent varies with the price profile (i.e., user classification is price profile dependent). In this model, a user's utility consists of a time-invariant function and a time-varying preference coefficient. This utility function assumes only diminishing returns of scale and continuous differentiability and can be quite general. Individual preferences over time are estimated by innovatively exploiting individual user volumes from the QPC experiment. (iii) The TDP design step formulates the pricing problem as a leader–follower game. The network manager maximises the bandwidth utilisation while keeping the total demand below the link capacity. Users maximise their own benefits under the given price profile and quota. The optimal price profile is then solved numerically. (iv) The network performance and user usage prediction step performs numerical extrapolation based on empirical data for effectiveness assessment. This step exploits the experimental data of step 1 and user demand model constructed by step 2 to simulate user behaviour. The network manager could then evaluate the network performance and user usage under the pricing policy designed by step 3. If the assessment is satisfactory, the network manager puts the pricing policy into practice. Otherwise, the network manager may tune the price difference between peak and off-peak hours by using heuristics such as incremental increase of peak-hour price or considering more pricing periods than just peak and off-peak hours.

3.1 User demand model under pricing

Although the concept of TDP is simple, the effective pricing design must be rooted in solid modelling of user demands

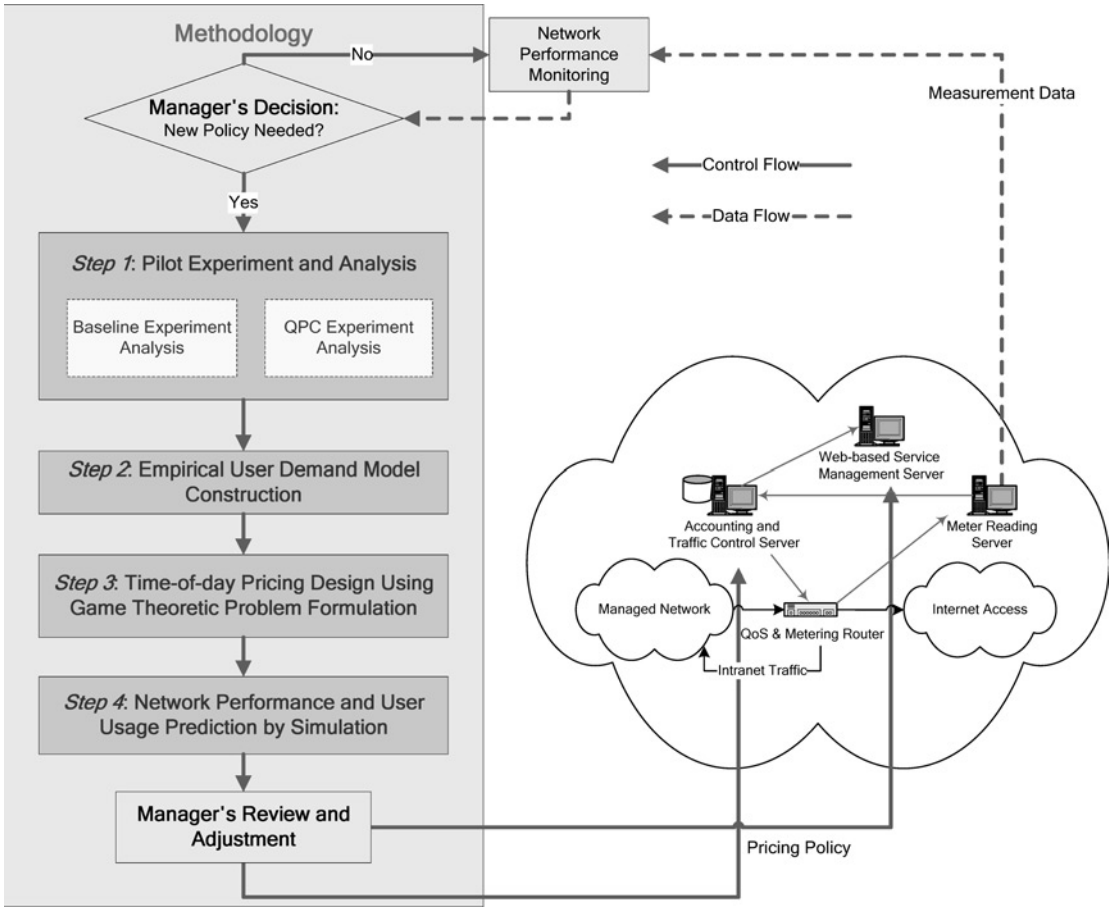


Fig. 4 *Flow-chart of design methodology of QPC/TDP*

and quota allocation behaviours. To mathematically model user behaviour under a given pricing policy, let us first define some notations.

Notations

B	bandwidth of Internet access;
T	length of a time slot;
$v_{i,k}$	Internet-access volume submitted by user i for regular service at time slot k, $i = 1, 2, \ldots, I$, and $k = 1, 2, \ldots, K$;
$v_{i,k,\mathrm{QPC}}$	Internet-access volume submitted by user i for regular service at time slot k in the QPC experiment, $i = 1, 2, \ldots, I$, and $k = 1, 2, \ldots, K$;
v_i^B	daily Internet-access demand of user i, obtained from the baseline experiment, $i = 1, 2, \ldots, I$;
Q	daily quota allotted to each user;
$Q_{i,k}$	remaining quota of user i at time slot k, $i = 1, 2, \ldots, I$, $k = 1, 2, \ldots, K$; note that $Q_{i,1} = Q$;
p_k	price (number of quota per byte) of regular service at time slot k, $k = 1, 2, \ldots, K$;
$\omega_{i,k}$	preference value of user i at time slot k, $i = 1, 2, \ldots, I$, $k = 1, 2, \ldots, K$;
S_j	set of type j users and $j \in \{m, p\}$, where m corresponds to the myopic type, while p is the prudent type.

3.1.1 User classification: Whether a user is myopic or prudent depends on the daily price profile announced by the network manager in advance and personal daily demand. If the daily Internet-access demand of a user can be satisfied at the maximal price, there is no need to be prudent in quota allocation. Such a user is therefore regarded as a myopic user. If a user's demand cannot be met at the minimal price, the user should be prudent and will carefully allocate the quota. For a user with the demand, which can be met at the minimal price but cannot be satisfied at the maximal price, the prudence of the user depends on a linear probability with respect to personal daily demand.

User classification algorithm

Input: v_i^B, which is measured from the baseline experiment, and $\boldsymbol{p} = [p_1 \quad p_2 \cdot p_k]$
Output: user type, S_j
For every user i, $i = 1, 2, \ldots, I$
If $v_i^B \le Q/\max_k\{p_k\}$, then $i \in S_m$,
else
if $Q/\max_k\{p_k\} < v_i^B \le Q/\min_k\{p_k\}$, then $i \in S_p$ with a probability x_i,

$$x_i \equiv \frac{v_i^B - Q/\max_k\{p_k\}}{Q/\min_k\{p_k\} - Q/\max_k\{p_k\}} \tag{1}$$

else
$i \in S_p$
end
end

3.1.2 Myopic user model: At time slot k, a myopic user i with $Q_{i,k} > 0$ determines the Internet-access volume of regular service, $v_{i,k}$, to maximise user is own benefit at that time slot only (short-term benefit), without considering

the value of $Q_{i,k}$ and the future demand. User i's benefit function (consumer surplus) [14] is

$$J_i^k(v_{i,k}) = U_i^k(v_{i,k}) - p_k v_{i,k}, \quad \text{for all } i \in S_m \tag{2}$$

where the first term represents the utility gain from submitting $v_{i,k}$ bytes and the second term represents the corresponding cost.

By taking the very common assumption in economics that users' utility follows diminishing returns to scale [14], we set

$$U_i^k(v_{i,k}) = \omega_{i,k} F(v_{i,k}) \tag{3}$$

where $F(v_{i,k})$ is strictly increasing, concave and continuously differentiable.

A user's quota accounting is based on the actually transmitted volume, which equals the submitted volume minus the dropped volume of a time slot. The drop ratio of regular service at time slot k is defined by

$$d_k \equiv \max\left[\frac{\sum_{i=1}^{I} v_{i,k} - \mathrm{BT}}{\sum_{i=1}^{I} v_{i,k}}, 0\right] \tag{4}$$

where BT is the capacity volume for Internet access over a time slot. Since users of regular class have the same transmission priority at the QoS router in Fig. 1, we assume that all regular-class traffic experiences the same drop ratio at this router. At the beginning of time slot $k+1$, user i's quota is therefore updated by

$$Q_{i,k+1} = Q_{i,k} - p_k v_{i,k}(1 - d_k) \tag{5}$$

Note that in our practical network architecture, the available quota will be updated every 10 min. Users can check their available quota at the web-based management server in Fig. 1.

Since what a myopic user i knows for decision at time k is whether the quota at that time is available or not, the myopic user decision problem (MUDP) is formulated as: (MUDP$_i$). For $k = 1, 2, \ldots, K$

$$\operatorname*{Max}_{v_{i,k}} U_i^k(v_{i,k}) - p_k v_{i,k}$$

subject to (4) and (5) with $Q_{i,k} > 0$.

Therefore when $Q_{i,k} > 0$, the optimal value, $v_{i,k}^*$, has to satisfy

$$\omega_{i,k} F'(v_{i,k}^*) = p_k \tag{6}$$

3.1.3 Prudent user model: At time slot k, a prudent user i considers the daily demand and allocates the available quota to determine the Internet-access volume of regular service for maximising user i's total benefit from time slot k to time slot K (long-term benefit), which is

$$\sum_{t=k}^{K} J_i^t(v_{i,t}) = \sum_{t=k}^{K} \left[U_i^t(v_{i,t}) - p_t v_{i,t}\right] \tag{7}$$

When planning for quota allocation, a user presumes that the planned submission would be transmitted, and the total submission should satisfy

$$\sum_{t=k}^{K} p_t v_{i,t} = Q_{i,k} \tag{8}$$

The prudent user decision problem (PUDP) is then (PUDP$_i$). For $k = 1, 2, \ldots, K$

$$\operatorname*{Max}_{\{v_{i,t}, t=k,\ldots,K\}} \sum_{t=k}^{K} \left[U_i^t(v_{i,t}) - p_t v_{i,t}\right]$$

subject to constraint (8) with $Q_{i,1} = Q$ and $Q_{i,k} > 0$.

Note that (PUDP$_i$) is a convex programming problem [15]. Consider the problem at time slot k with $Q_{i,k} > 0$. To develop the optimality conditions, form a Lagrangian function first

$$L_i(v_{i,t}, t = k, \ldots, K; \lambda_i) = \sum_{t=k}^{K} \left[U_i^t(v_{i,t}) - p_t v_{i,t}\right] - \lambda_i \left(\sum_{t=k}^{K} p_t v_{i,t} - Q_{i,k}\right)$$

where λ_i is a Lagrangian multiplier. The optimality conditions are then

$$\frac{\partial L_i}{\partial v_{i,t}} = \omega_{i,t} F'(v_{i,t}) - (\lambda_i + 1) p_t = 0, \quad t = k, \ldots, K \tag{9}$$

and

$$\frac{\partial L_i}{\partial \lambda_i} = \sum_{t=k}^{K} p_t v_{i,t} - Q_{i,k} = 0 \tag{10}$$

Based on (9) and (10), it can be shown that the optimal value, $v_{i,t}^*$, must satisfy

$$\omega_{i,t} F\prime(v_{i,t}^*) = \frac{p_t}{Q_{i,k}} \sum_{j=k}^{K} \left[\omega_{i,j} F\prime(v_{i,j}^*) v_{i,j}^*\right], \; t = k, \ldots, K \tag{11}$$

3.1.4 Estimation of user preferences: To estimate $\{\omega_{i,k}\}$, we ingeniously exploit the measured data (the submitted volume) by conducting a QPC experiment, which means QPC/TDP with $p_k = 1$, $k = 1, 2, \ldots, K$. Given a function $F(v_{i,k})$ in (3). Let $v_{i,k,\mathrm{QPC}}$ be the submitted volume of user i for regular service at time slot k in the QPC experiment. For a myopic user, the preference value is $\omega_{i,k} = p_k / F'(v_{i,k})$ based on (6). Given (6) and $v_{i,k,\mathrm{QPC}}$ with $p_k = 1$, $k = 1, 2, \ldots, K$, user i's preference at time slot k is estimated as

$$\omega_{i,k} = 1 / F'(v_{i,k})|_{v_{i,k} = v_{i,k,\mathrm{QPC}}} \tag{12}$$

For a prudent user, we further assume that the user allocates one's quota only at the time of quota replenishment, that is, $k = 1$. Let $X_t = \omega_{i,t} F'(v_{i,t}^*)$, $t = 1, \ldots, K$ and substitute them into (11) to solve X_t, $t = 1, \ldots, K$. Solving the linear equations, we obtain $X_t = C p_t$, $t = 1, \ldots, K$, where C is an arbitrary constant. Without loss of generality, let $C = 1$ and the preference value of a prudent user i is then estimated as $\omega_{i,t} = p_t / F'(v_{i,t})$, $t = 1, \ldots, K$. Therefore given the submitted volume in the QPC experiment, the preference of a prudent user can be estimated by (12) as well.

3.1.5 Utility function selection: Our demand models require the utility function, which satisfies continuous differentiability and diminishing returns to scale. To convey further discussions, let us now select a specific form of $F(v_{i,t})$ without loss of generality of the methodology. Researchers of [16, 17] modelled the utility as a logarithmic function of the rate based on the property of diminishing marginal utility. Since the submitted volume of a time slot

can be approximated by a constant submission rate multiplied by the slot duration T, we also model a user's utility as a logarithmic function of the submitted volume that is

$$F(v_{i,k}) = \log v_{i,k} \tag{13}$$

Given (13), the solutions to (MUDP$_i$) and (PUDP$_i$) are then

$$v_{i,k}^* = \frac{\omega_{i,k}}{p_k} \tag{14}$$

and

$$v_{i,t}^* = \frac{\omega_{i,t}}{\sum_{j=k}^{K} \omega_{i,j}} \frac{Q_{i,t}}{p_t}, \quad t = k, \ldots, K \tag{15}$$

respectively, with preference value

$$\omega_{i,k} = v_{i,k,\mathrm{QPC}} \tag{16}$$

3.2 Pricing problem design

As the service charge is flat-rate or free, the goal of price setting by a network service provider (NSP) is to maximise the total bandwidth utilisation over a day while the average total demand does not exceed the capacity. Mathematically, the bandwidth utilisation of regular service at time slot k is defined as

$$\frac{1}{\mathrm{BT}} \sum_{i \in A_k} v_{i,k}(1 - d_k) \tag{17}$$

where $A_k = \{i | Q_{i,k} > 0, \ \forall i\}$ is the set of users whose available quota at time slot k is non-zero, and $\sum_{i \in A_k} v_{i,k}(1 - d_k)$ represents the total transmitted volume of regular service at time slot k.

The constraint that the total expected volume of submission cannot exceed the link capacity at a time slot is expressed as

$$\sum_{i \in A_k} v_{i,k} \le \mathrm{BT}, \quad k = 1, 2, \ldots, K \tag{18}$$

As for network management in practice, the NSP may allow a safe margin of bandwidth B_{rev} to prevent the network from instability. The constraint (18) can be modified as $\sum_{i \in A_k} v_{i,k} \le (B - B_{\mathrm{rev}})T$, $k = 1, 2, \ldots, K$.

Taking users' behaviours characterised by (MUDP$_i$) and (PUDP$_i$) into consideration, the NSP has a pricing problem (PP), formulated as

$$\text{(PP)} \quad \underset{\{p_k \in \Omega, \ k=1,2,\ldots,K\}}{\mathrm{Max}} \frac{1}{\mathrm{BT}} \sum_{k=1}^{K} \sum_{i \in A_k} v_{i,k}(1 - d_k)$$

subject to constraints (4), (18), the user classification algorithm and MUDP$_i$ (PUDP$_i$) if user i is myopic (prudent).

The pricing problem is formulated as a Stackelberg game [13], where the NSP acts as the leader and users are followers. The optimisation of the TDP profile is nested with individual users' optimisation of submitted volumes. Although the solutions to individual user's optimisation problems have closed forms, the closed-form analytic solution of (PP) may not be available because the objective function and constraint (18) are nonlinear. The numerical method is thus adopted to solve (PP). The optimal price profile of (PP) is obtained by exhaustive search over the feasible price profiles. Given each leader's admissible price profile, we solve individual followers' optimisation problems [(6) and (9)–(11)], and substitute the user submitted volume into the leader's objective and obtain the corresponding bandwidth utilisation. Among the enumerations of all the leader's price profiles, we select the one that induces a maximum of bandwidth utilisation while meeting the link capacity constraint.

4 Pricing policy over simple examples

Simple examples are designed to study (i) how the price induces behaviours of myopic and prudent users, respectively, and thus affects the network performance, and (ii) how the TDP policy varies with respect to user behaviours. Consider the regular service of a network with a bandwidth of ten units. There are three users and three time slots. Individual user preferences in Table 1 monotonically increase with time for easy analysis of the relationship between price and preference. There are five feasible price values, $\Omega = \{1, 2, 3, 4, 5\}$, whose range is large enough that there exists a feasible price profile to keep the total user demand within the capacity.

4.1 Pricing policy for prudent users

It is intuitively clear that because of the limited bandwidth, the higher the user preference value, the higher the price for a time slot that is $p_1^* \le p_2^* \le p_3^*$. Such an intuition can actually be justified when all users are prudent. According to (15) and to satisfy the link capacity constraint that is

$$\sum_{i=1}^{I} v_{i,k}^* = \sum_{i=1}^{I} \left(\frac{\omega_{i,k}}{\sum_{t=k}^{K} \omega_{i,t}} \frac{Q_{i,k}}{p_k} \right) \le \mathrm{BT}$$

p_k must satisfy

$$p_k \ge \sum_{i=1}^{I} \frac{W_{i,k} Q_{i,k}}{\mathrm{BT}} \tag{19}$$

where $W_{i,k} = \omega_{i,k} / \sum_{t=k}^{K} \omega_{i,t}$. For maximisation of bandwidth utilisation, p_k^* must be as low as possible because the submitted volume is proportional to the reciprocal of the price by (15). So

$$p_k^* = \sum_{i=1}^{I} \frac{W_{i,k} Q_{i,k}}{\mathrm{BT}} \tag{20}$$

Note that $p_1^* \le p_2^* \le p_3^*$ because $\{W_{i,k}\}$ in (20) is monotonically increasing with k in this example. This property among p_k's holds for any given quota value $Q > 0$.

When $Q = 10$ and $Q = 25$, $(p_1^*, p_2^*, p_3^*) = (1, 1, 2)$ and $(p_1^*, p_2^*, p_3^*) = (2, 3, 4)$, respectively. They indeed have the expected property. In Table 2, the total submitted volume of time slot 3 is reduced by 50% (peak shaving) as compared with QPC, alleviating the congestion.

Table 1: User preference profiles

	Time slot 1	Time slot 2	Time slot 3
User 1	7	9	11
User 2	5	7	9
User 3	3	5	7

Table 2: Total submitted volumes of QPC and QPC/TDP in the case of prudent users when $Q = 10$

	Time slot 1	Time slot 2	Time slot 3
QPC scheme	6.97	10	13.03
QPC/TDP scheme	6.97	10	6.51

4.2 Pricing policy for myopic users

The property that $p_1^* \le p_2^* \le p_3^*$ no longer holds when users are all myopic; it depends on the quota value. A myopic user i's submitted volume at time slot k is given in (14). When there exists a sufficient quota, the total desired volume is $\sum_{i \in A_k} \omega_{i,k}/p_k$. Constrained by the link capacity,

$$p_k \ge \sum_{i \in A_k} \frac{\omega_{i,k}}{\mathrm{BT}} \tag{21}$$

From the NSP's point of view, one byte transmitted at one time slot is of the same contribution to bandwidth utilisation as one byte in any other time slot. So, the NSP would select p_k^* to be as low as possible, namely

$$p_k^* = \sum_{i \in A_k} \frac{\omega_{i,k}}{\mathrm{BT}} \tag{22}$$

Based on such a basic understanding, let us examine the pricing policies when $Q = 10$ and $Q = 25$. When $Q = 10$, $(p_1^*, p_2^*, p_3^*) = (2, 3, 1)$, the property that $p_1^* \le p_2^* \le p_3^*$ does not hold. Table 3 indicates that the average submitted volume is reduced by 66% as compared to QPC and the total submitted volume across all time slots is around 7 (load balancing). When $Q = 25$, $(p_1^*, p_2^*, p_3^*) = (2, 3, 3)$, where monotonicity holds, as in the case of prudent users only.

4.2.1 Remarks: Our methodology is generic enough to handle problems with different objectives as long as individual users' utility satisfies diminishing returns to scale. For example, the methodology can be directly applied to the case where the NSP's objective function is the social welfare function $\sum_{i=1}^{I} \sum_{k=1}^{K} [U_i^k(v_{i,k}) - p_k v_{i,k}]$, which is commonly used in economics. Solution properties are analysed as follows. At time slot k, user i's benefit, $U_i^k(v_{i,k}) - p_k v_{i,k}$, is proportional to $-\log p_k$ according to (13)–(15). To maximise the social welfare (summation of all users' benefits), p_k should be as low as possible under the link capacity constraint. The resultant prices and control effects are identical to Sections 4.1 and 4.2.

4.3 Importance of user differentiation on time-of-day management

Table 4 shows the patterns of the total submitted volume under the biased pricing policies. Case I (II) means that the pricing policy is designed for the prudent (myopic) behaviour, but all users are myopic (prudent). In Case I, the submitted volumes across all time slots are not shaved, higher than 10. Case II shows that bandwidth utilisation is less than 50% at time slots 1 and 2, but the congestion happens at time slot 3. These results reveal that if the pricing scheme is designed without differentiating the behaviours between myopic and prudent users, it easily leads to either serious congestion in peak hours or bandwidth waste in off-peak hours.

Table 3: Total submitted volumes of QPC and QPC/TDP in the case of myopic users under $Q = 10$

	Time slot 1	Time slot 2	Time slot 3
QPC scheme	15	21	27
QPC/TDP scheme	7.5	7	7

Table 4: Total submitted volumes in cases I and II under $Q = 10$

	Time slot 1	Time slot 2	Time slot 3
Case I	15	21	13.5
Case II	3.49	3.33	13.03

5 Effectiveness assessment

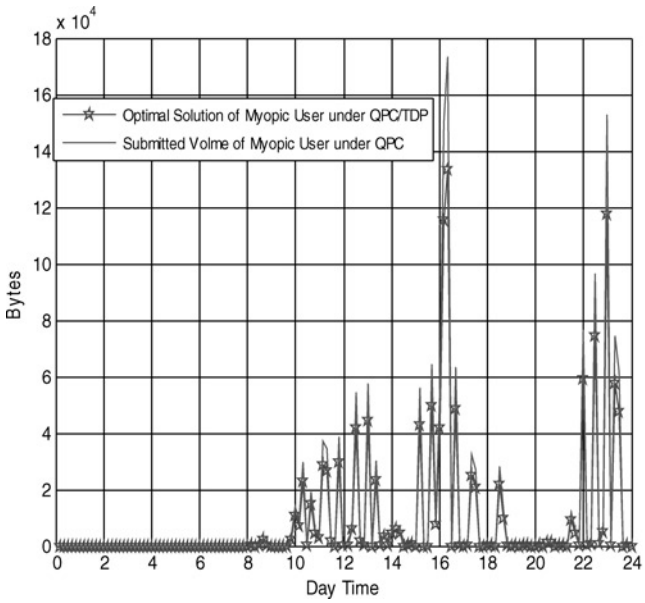
This section applies step 4 of the QPC/TDP design methodology and exploits the empirical data of NTU dormitory networks to set up a numerical experiment for investigating control effects of QPC/TDP on congestion reduction, and uneven usage and fairness improvement. In this step, the daily quota for each user is set to 1 GB. We utilise the QPC experimental data of step 1 to estimate user preferences over time by (16) derived from step 2. The empirical data of the baseline experiment in step 1 serve for user classification by (1). The 54 Mbps bandwidth for outbound dormitory traffic is the bottleneck. Quota replenishment takes place at 6 a.m. every day. As the NTU network management system collects metering data once every 10 min, accordingly, the length of a time slot is set to 10 min. There are six levels of price per byte of transmission, $\Omega = \{0.9, 1, 1.1, 1.2, 1.3, 1.4\}$. Such a price range is large enough so that there exists a feasible price profile to make the total user demand within the link capacity.

In the experiment, we hypothesise that $p_{\text{off-peak}}^* < p_{\text{peak}}^*$ and the drop rate of regular service equals 0, as designed. During peak hours, both the total submitted volume of regular service and user transmitted volume of Internet access significantly decrease over pure QPC because the peak-hour price curbs usage.

Under the game-theoretic formulation, numerical results show that the optimal price profile is $p_{\text{off-peak}}^* < p_{\text{peak}}^* = (1.1, 1.3)$, where peak hours are from 9 a.m. to 3 a.m. The optimal solutions of two selected users are depicted in Fig. 5, where one is myopic and the other is prudent. They indicate that the submitted volumes (optimal solutions) of two users under QPC/TDP are significantly reduced by about 17% during peak hours as compared to those under QPC only.

5.1 Peak shaving, load balancing and congestion alleviation

Table 5 shows that the total submitted volume of regular service under QPC/TDP is 11.53% less than that of the QPC scheme during peak hours. This peak shaving effect is a result predicted by (14) and (15), as the submitted volume of regular service is proportional to $1/p_{\text{peak}}^*$, where $p_{\text{peak}}^* > 1$. In Fig. 6, the difference of the average submission between peak and off-peak hours is reduced by 31.21%, namely a load'balancing effect is achieved. In addition, the drop rate is reduced from a maximum of 5 Mbps under QPC to zero at all times by the price control.



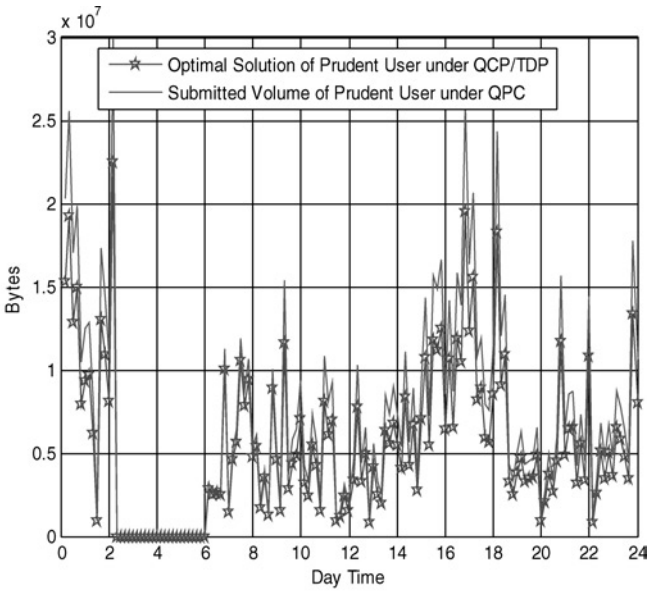


Fig. 5 *Optimal solutions of users under QPC/TDP*

5.2 Usage reduction of heavy users in peak hours

Table 6 shows the reduction percentage of Internet access over time by the top five users under QPC/TDP. During peak hours, there is an average of 17.62% reduction from that of QPC, and 4.4% from that of the 500 MB ∗ 2 QPC scheme, where the 500 MB ∗ 2 QPC scheme means that 500 MB quotas are allotted at 6 a.m. and 6 p.m., respectively, with no quota carrying over to the next time period. At individual slots, there is an average of 15.46% reduction as compared with QPC. Fig. 7 reveals that Internet-access usage by the top five users under QPC/TDP is less than that of QPC all the time.

5.3 Fairness improvement in peak hours

The standard deviation of total users' usage for Internet access is chosen as a metric of fairness. Again, the performance of QPC serves as the basis of comparison unless specified. In Table 7, there is an average of 17.64% reduction during peak hours. There is an average of 8% reduction as compared with the 500 MB*2 QPC scheme. Since $v_{i,k}$ in peak hours is proportional to $1/p^*_{\text{peak}}$, the standard deviation of the submitted volume for regular service is proportional to $1/p^*_{\text{peak}}$, which will be reduced with the increase of p^*_{peak}. For individual time slots, there is an average of 15.44% reduction. The standard deviation of total users' usage over time is depicted in Fig. 8, where the standard deviation under QPC/TDP is reduced across all the time slots, especially in peak hours.

Table 5: Average total submission rate and drop rate under QPC and QPC/TDP schemes during peak hours

	QPC scheme	QPC/TDP scheme
Average total submission rate (Mbps)	58.03	51.35
Average drop rate (Mbps)	4.03	0

5.4 Policy adaptation to changes

For price adaptation to network changes, the baseline and QPC experiments have to be conducted for a short period, say one week [1], respectively, to determine the new user characteristics. For example to manage the NTU dormitory network, the administration usually needs to conduct the baseline and QPC experiments only at the beginning of a new academic year, when new residents move in. Note that the baseline experiment is a special case of QPC/TDP, where all prices are set to 1 and the daily quota is infinite, and that the QPC experiment is also a special case of QPC/TDP, where all prices are set to 1 but with a finite quota. In the case of the NTU dormitory network with 5355 users, the policy design and assessment take only 1 min (CPU time) in a Matlab^{TM} environment, where it runs on a PC with an Intel Pentium M processor and 512 MB memory. If the network manager sets three level prices, p_{H}, p_{M} and p_{L}, $(p^*_{\text{H}}, p^*_{\text{M}}, p^*_{\text{L}}) = (1.3, 1.1, 0.9)$, where p_{H} corresponds to the price between 9 a.m. and 3 a.m., p_{M} corresponds to the prices between 3 a.m. to 5 a.m. and 7 a.m. to 9 a.m. and p_{L} is the price between 5 a.m. and 7

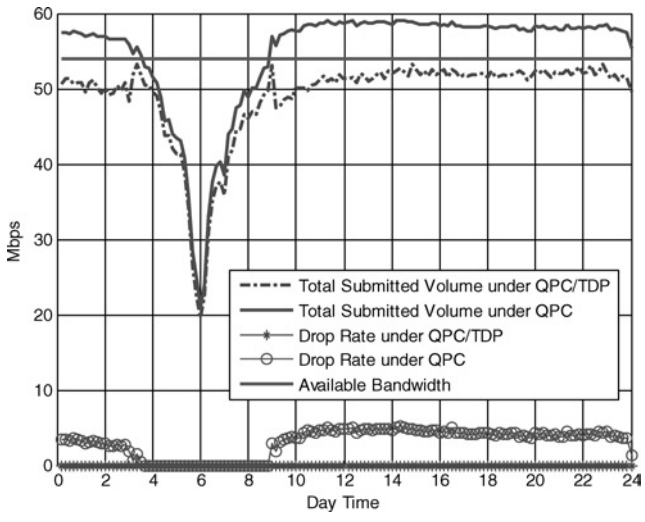


Fig. 6 *Total submitted volumes and drop rates of regular service under QPC and QPC/TDP schemes*

Table 6: Reduction percentage of internet-access usage by top five users under QPC/TDP

	QPC scheme	500 MB * 2 QPC scheme
Individual time slot	−15.46%	−0.3%
Peak-hour time slot	−17.62%	−4.4%

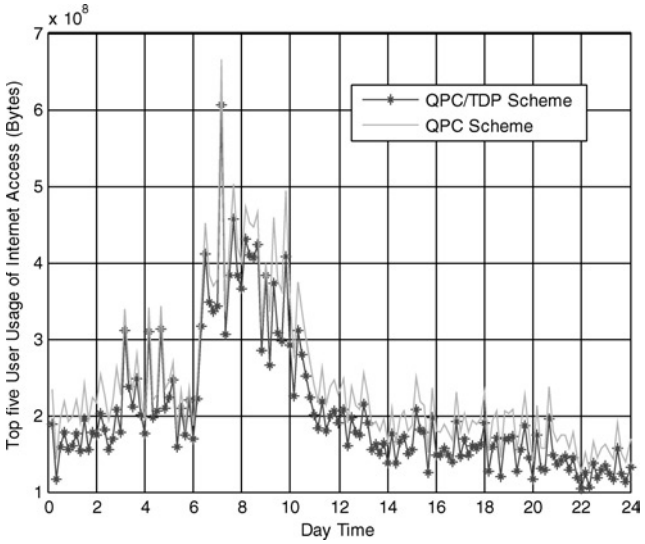


Fig. 7 *Top five user internet-access volumes under QPC and QPC/TDP schemes*

a.m. Since $p_L^* < p_{\text{off-peak}}^*$, user usage during 5 a.m. to 7 a.m. is more encouraged as compared to two-level prices. The computation time in this case takes less than 7 min. With a short time period for data collection and fast policy design and evaluation, the design methodology could respond to a changing network environment.

5.4.1 Remarks: There is a difference between the average user usage predicted by our models and the real-time (or instantaneous) user usage because our models average across users and over time periods (10 min). The models are adopted to help the network manager with pricing design for managing network traffic in average. When the observed average user submission volumes significantly deviate from those the mathematical models are based on, a network manager may consider conducting the pilot experiments to update user behaviour model. Then, the manager follows the methodology described in Section 3 to adjust QPC/TDP policy parameters accordingly.

5.5 Comparison with existing management schemes

Many network administrations [18, 19] adopt quota-based control schemes to regulate users' network service access. Some even curb users' usage by applications when a user's usage exceeds the given quota. The user will be prohibited from accessing the Internet for a time period based on the excess volume. Such a scheme is compulsive and does not take users' temporal usage pattern into account. Our proposed QPC/TDP scheme not only allows the users who run out of the quota to access the Internet with a lower priority but also provides managers with a time-of-day traffic management function.

Table 7: Reduction percentage of the standard deviation for internet-access usage under QPC/TDP

	QPC scheme	500 MB * 2 QPC scheme
Individual time slot	−15.44%	−0.32%
Peak-hour time slot	−17.64%	−8%

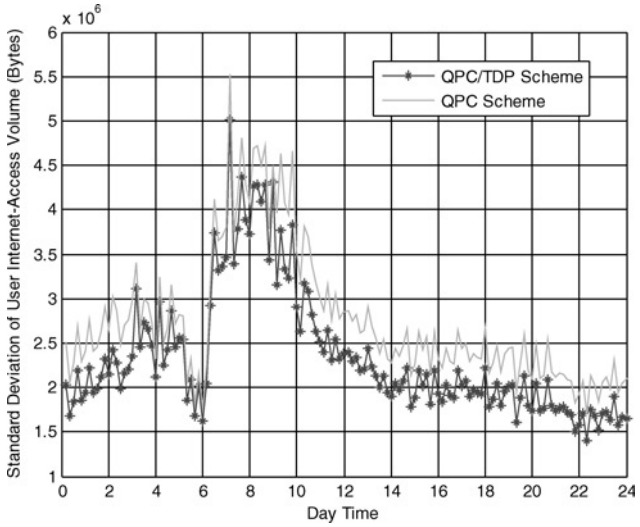


Fig. 8 *Standard deviations of user Internet-access volume under QPC and QPC/TDP schemes*

6 Conclusions

This paper considers the problems of congestion, uneven usage and unfairness during peak hours over a free-of-charge or flat-rate network. The five specific challenges presented in Section 2 are all addressed. We grasp user behaviours of quota allocation observed from empirical data and construct myopic and prudent user models, answering (C1). The price design to maximise bandwidth utilisation is obtained by solving a leader–follower game, addressing (C2). For (C3) and (C4), we adopt static TDP because it is not only simple and acceptable for users but also easy to implement over a QPC environment. The design methodology of QPC/TDP is proposed to address (C5). Evaluation of our methodology shows that the peak-hour usage by heavy users and fairness are improved by 17% as compared to QPC only. The average total submission during peak hours is reduced by 12%, and the difference between peak and off-peak hours is reduced by 31%. The design methodology itself requires only baseline and QPC experiments for one week each and is efficient for pricing calculation. It can apply to frequently changing Internet-access environments such as campus, government, community and corporate LANs.

7 Acknowledgment

This work was partly supported by the National Science Council, Taiwan, Republic of China, under grants NSC-91-2219-E-002-033, NSC-92-2212-E-002-060 and NSC-93-2212-E-002-082.

8 References

1 Lin, T.-C., Sun, Y.S., Chang, S.-C., Chu, S.-I., Chou, Y.-T., and Li, M.-W.: 'Management of abusive and unfair internet access by quota-based priority control', *Comput. Netw.*, 2004, **44**, pp. 441–462

2 Mitchell, B.M., and Vogelsang, I.: 'Telecommunications pricing: theory and practice' (Cambridge University Press, 1991)
3 Farmer, E.D., Cory, B.J., and Perera, B.L.P.P.: 'Optimal pricing of transmission and distribution services in electricity supply', *IEE Proc.,-Gener., Transm. Distrib.*, 1995, **142**, (1), pp. 1–8
4 Pillai N.V.: 'A contribution to peak load pricing theory and application', 2003, available at http://ideas.repec.org/p/ind/cdswpp/346.html
5 Brunekreeft, G.: 'Price capping and peak-load pricing in network industries', 2000, available at http://www.vwl.uni-freiburg.de/fakultaet/vw/publikationen/diskussionspapiere/disk73.pdf
6 Fitkov-Norris, E.D., and Khanifar, A.: 'Dynamic pricing in mobile communication systems'. Proc. First Int. Conf. on 3G Mobile Communication Technologies (IEE Conf.), 2000, pp. 416–420
7 Carroll, J.E., and Kirkby, P.A.: 'Proportionally fair pricing: dynamics, stability and pathology', *IEE Proc., Commun.*, 2000, **147**, (1), pp. 23–31
8 Mackie-Mason, J.K., and Varian, H.R.: 'Pricing the internet' in Kahin, B., and Keller, J. (Eds.): 'Public access to the internet' (Prentice-Hall, 1994)
9 Jin, N., Venkitachalam, G., and Jordan, S.: 'Dynamic pricing of network resources', Proc., IEEE GLOBECOM'03, 2003, **6**, pp. 3216–3220
10 Paschalidis, Ch., and Tsitsiklis, J.N.: 'Congestion-dependent pricing of network services', *IEEE/ACM Trans. Netw.*, 2000, **8**, (2), pp. 171–184
11 Shih, J.S., Katz, R.H., and Joseph, A.D.: 'Pricing experiments for a computer-telephony-service usage allocation', Proc., IEEE GLOBECOM'01, 2001, **4**, pp. 2450–2454
12 Edell, R., and Varaiya, P.: 'Providing internet access: what we learn from INDEX', *IEEE Netw.*, 1999, **13**, (5), pp. 18–25
13 Osborne, M.J.: 'An Introduction to Game Theory' (Oxford University Press, 2003)
14 Parkin, M.: 'Economics' (Addison Wesley, 1994)
15 Minoux, M.: 'Mathematical programming: theory and algorithms' (Wiley, Chichester, 1986)
16 Kelly, F.P.: 'Charging and rate control for elastic traffic', *Eur. Trans. Telecom*, 1997, **8**, pp. 33–37
17 Wang, X., and Schulzrinne, H.: 'Pricing network resources for adaptive applications in a differentiated services network', Proc., IEEE INFOCOM'01, 2001, **2**, pp. 943–952
18 Available at http://net.nthu.edu.tw/
19 Available at http://www.cc.nccu.edu.tw/network/dormnet/