Fair and Performance Guaranteed Methods for Flat-Rate Unlimited Access Service Plan

Yeali S. Sun[1], Pei-Wen Chen[1], Meng Chang Chen[2]

[1]Dept. of Information Management, National Taiwan University
sunny@im.ntu.edu.tw
[2]Institute of Information Science, Academia Sinica, Taiwan
mcc@iis.sinica.edu.tw

Abstract. Simplicity in administration and operation is the choice for production networks. Hence, flat-rate unlimited access service plan is the predominant form of retail pricing in Broadband Internet access services. However, this service plan can easily result in unfair resource sharing, abusive usage and poor performance. In this paper, we propose several fair and performance guaranteed methods to alleviate the problems. The working field trial project called Virtual Internet Pricing (VIP) project was deployed at the dormitory network of National Taiwan University with a total of 5355 users. In VIP, a quota-based priority control (QPC) scheme was proposed to resolve the problems. While it alleviated the problems, QPC however raised some issues such as chaotic periods, bandwidth stealing and weak performance guarantees for in-profile packets. Four methods were proposed to enhance the basic QPC scheme. The simulation results showed that the proposed methods significantly improved network performance and increased system stability.

1 Introduction

In the past, a number of Internet pricing models (such as [1,2,3,4]) have been proposed. The common argument is in favor of the model of charging users by actual usage as in many public utility services. Despite of the fact, today's Internet access services (e.g., dial-up, ADSL and cable modem) are mostly charged by flat rates probably for different access speeds or a fixed fee plus per unit time charge [5,6]. As pointed out in [6], the flat-rate unlimited-access service plan is economically inefficient. Because users do not face the true marginal cost of usage, it often results in over-usage, jeopardizing network performance. Studies also find that there is big usage difference of users between being served under the flat-rate unlimited access plan and being served under the usage-based charge plan. Hence, many service providers and network operators take the tactic of 50% or 70% threshold utilization as the rule of thumb in upgrading their resources/systems to avoid congestion.

We faced similar problems as do commercial networks in our university dormitory network. There were twelve dormitories with a total of 5355 students who paid a one-time network access fee for every semester. Each student was given a fixed IP address and a Fast Ethernet access to the Internet. The *unlimited* access service

caused slow Internet access and frequent connection timeouts. After preliminary investigation and analysis, we found less than 10% of the total users contribute more than 90% of the daily traffic. When looking into individual user traffic, we found that this small group of *"heavy"* users had voluntarily contributed their computers to form various kinds of peer-to-peer networks for file downloading and content sharing. While large files were downloaded from these machines, largely 90% "regular" users in the dormitory network follow the typical internet access pattern.

Essentially, the problem we encountered is that, under the flat-rate unlimited service plan, users are not charged on the basis of how many packets are sent. Without control, user usage could become excessive and outrageous, causing severe network congestion, performance degradation and unfair resource sharing. Since network is not free, light users are indeed subsidizing heavy users. Similar problem also exist in today's intranets such as schools, companies, commercial buildings and residential community that share common links to the Internet. Those intranets are generally implemented with high-speed technology such as Gigabit/Fast Ethernet. But the bandwidth of the Internet access links are however much lower. Many of these network users experience congestion on the Internet access link. In some cases, the congestion period could last for almost entire day.

To address the poor Internet access performance and unfair resource sharing caused by excessive (selfish) use of a small number of users, we conducted an experimental field trial project called Virtual Internet Pricing (VIP) [7]. We consider performance incentive as an alternative to address the fairness and performance problems that result from the flat-rate unlimited access service plan. Although per-flow scheduling like Weighted Fair Queueing algorithm is widely considered as a good technique to enforce fairness and QoS guarantees in a link-sharing environment[8,9], not many network equipment implemented per-flow Quality of Service (QoS) due to its overhead. In a *Quota-based Priority Control* (QPC) system, each user is allowed to transmit no more than a maximum amount of high-priority traffic in each quota control period. (Note those packets are called *in-profile* packets.) We adopted QPC scheme to enforce per-user fairness and to relieve the congestion problem caused by a small group of selfish heavy users.

The QPC scheme however raises several interesting problems including chaotic periods, bandwidth stealing and weak performance guarantees for in-profile packets. When many users are backlogged at the beginning of the quota control period, undesirable congestion often occurs as soon as the period begins. Packet loss periods could last longer than thirty minutes with a loss rate of more than 10%. This phenomenon is referred to as the *chaotic period*.

Traffic metering is typically implemented in routers and the accounting is performed at the backend systems [11]. Accounting systems periodically collect user usage data from the metering routers. The accounting data collection interval is an important system parameter in quota control. To minimize the overhead, accounting interval is often set to the value much greater than packet transmission time (e.g., every 10 minutes). This way however makes the *bandwidth-stealing* possible. Because per-user account balance is checked per accounting interval, during the period a user may over use or steal the bandwidth. It directly affects the effectiveness of the

QPC scheme in enforcing fairness and the performance guarantees to in-profile packets.

In QPC there are ways to overrun a bottleneck resource. For example, a user with a full quota account balance can legally introduce extremely bursty in-profile packets into the network within a short period of time. This makes the provision of performance guarantees to in-profile packets quite challenging. We will examine and discuss each of these problems in detail later in this paper.

1.1 Related Work on Internet Pricing

Works have been proposed to devise optimal pricing policy for optimal social welfare. Achieving this, a marginal congestion cost is charged. Congestion costs are the performance penalties incurred from imposing one user's traffic on other users. There are many ways to deal with congestion externalities, such as to establish social norms, to establish a rationing or quota system or to develop a pricing mechanism. In the literature, a group of researchers especially economists prefer using pricing to manage user behavior in dealing with the problems of network resource scarceness and congestion. The advantage of this approach is that one can effectively control network traffic as well as achieve economic profits. A good example is Shadow price [3]. However, there are three possible issues as pointed out in [4]. First, marginal congestion cost-based pricing may not produce sufficient revenue to fully recover actual costs. Second, congestion costs are difficult to characterize and obtain from the network, and therefore cannot reliably form the basis for pricing. Third, there are other more structural goals besides optimality.

Other works on pricing include smart market [2] and edge pricing [4]. In edge pricing, true congestion costs are approximated by replacing actual congestion conditions and the cost of actual paths with expected congestion conditions and the cost of expected paths. Under this model, charges depend only on source and destination pairs and therefore can be determined and assessed locally at the access point rather than computed in a distributed fashion along the entire path.

2. The VIP (Virtual Internet Pricing) Project

In addition to the unfairness and congestion problems, the university campus network administrators also face the problem that the total traffic emitted from the dormitory network constitutes more than 50% of the total traffic in then the ATM 155Mbps campus backbone. Initially, the university network administrators decided to impose a maximum rate of 54Mbps on the traffic from the dormitory networks. Unfortunately, the uplink became even more congested. As a result, the VIP project was initiated and an experimental field trial of virtual internet pricing was conducted in the NTU dormitory network to solve the problem. The traffic-metering device is a Cisco router in which Netflow collects IP usage data. The QoS router is a home-made Linux-based device implementing priority scheduling and IP packet classification. The meter-reading server performs usage data analysis. Per-user usage accounting, service charging are implemented on a server machine which is also responsible

for sending traffic control commands to the QoS router to real-time configure individual user's priority levels in accordance with the account balance.

In the Virtual Internet Pricing (VIP) project, the dormitory network severed 5355 users. For the purpose of administrative and operational simplicity, the quota-based control combined with priority scheduling was used to achieve per-user fairness under the flat-rate unlimited service plan. There are two services: the *Regular* service and the *Custody* service, corresponding to the high and low priority queues, respectively. The default service is high-priority regular service. Each user is given an *account* of *virtual network dollars* (abbreviated as Net$). At the beginning of a *quota control period*, each account is credited with a fixed amount of Net$. Traffic sent through different priority queues is charged at different rates (e.g., Net$10/Mbits for high priority and no charge for low priority). During the period, the system records the amount of traffic (in bytes) each user transmits over the uplink and accordingly deducts service charges from the account. When a user's account balance becomes zero, the user is classified as *under custody* – subsequent traffic from the user is marked as low priority until the current quota control period ends.

There are two performance goals. The first goal is to satisfy the bandwidth demands of the majority users under max-min fairness. Assume users are numbered from 1 to N according to the increasing order of their traffic demands. To satisfy the demands of the first *s* of the total user population, the quota is set to be equal to or more than the demand of the s^{th} user. Note that this quota assignment guarantees the minimum bandwidth to every user. During a quota control period, in-profile traffic is transmitted as high priority. Excess traffic will be marked as out-profile packets, served by using the remaining bandwidth in the low-priority queue.

The second goal is to guarantee a maximum average packet loss rate for in-profile packets in each quota control period. Given the user population and bottleneck link capacity, the performance of the high priority queue directly relates to the quota assignment. The larger the quota, the more the traffic ranked as high priority, and possibly, the greater packet loss in the queue. If, however, the quota amount is too small, most users' basic traffic demands will not be met.

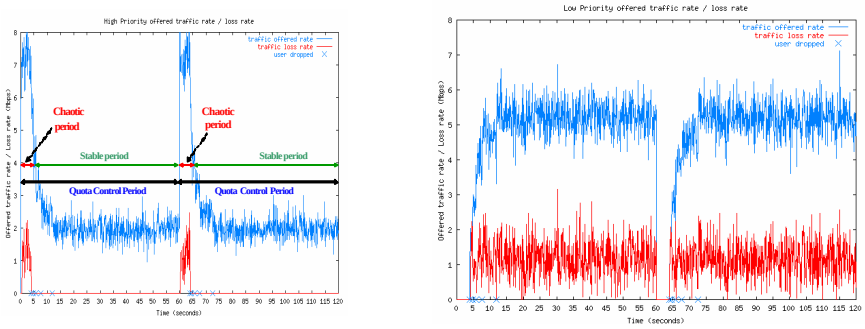
3. Problems in Quota-based Priority Control

Three problems – chaotic periods, bandwidth stealing and weak performance guarantees to in-profile packets obstructed the two performance goals.

A. Chaotic Periods

In the field trial, congestion occurred almost at the beginning of every quota control period. Because every user's account was deposited with full credits at the beginning of a control period, suddenly bursts of packets were injected into the high-priority queue and the queue quickly built up. The situation usually continued for half an hour and resulted in severe loss of in-profile packets until heavy users gradually used up quota. This congestion period is referred to as the chaotic period. We presented the

simulation results in Figure 1 o illustrate the traffic load and packet loss rate dynamics during a chaotic period. In the simulation, the link capacity is 6Mbps, the quota amount is 6Mbits, the quota control period is 60 seconds, the accounting interval is per-packet, and there are 100 users. In the Figure, the chaotic period lasted for about 13 seconds until a sufficient number of users used up quotas and their packets were thereafter directed to the low-priority queue. After then, the high priority queue becomes stable and the performance of the in-profile packets was no longer affected by the out-profile packets in the low-priority queue. Besides quota-based control can achieve long-term fairness, it leaves users great flexibility in sending packets. Even we interleave user with different starting time of quota control period, chaotic period might still happens when some users transmit in-profile packets at very high transmission rate. In practice, due to the dynamics of user demands and behavior, it is difficult to predict when a chaotic period will take place and how long it will last. It therefore poses great challenges in resource planning to support performance guarantees such as maximum packet loss rate to in-profile packets under the basic QPC scheme.



(a) Regular service (high priority queue).

(b) Custody service (low priority queue).

Fig. 1 Severe packet loss in the high-priority queue during a chaotic period.

B. Bandwidth Stealing

By taking into consideration the operational issues, such as feasibility, complexity and overhead, of this large-scale field trial, we decided to perform traffic data collection and usage accounting every 10 minutes. During the accounting interval, a user may generate more in-profile packets than allowed, i.e. over-use. Such *bandwidth-stealing* from heavy users happened in the field trial. It unfortunately worsened the congestion situation in the chaotic periods. It also degraded fairness in the quota-based control. A possible amendment is to reduce accounting interval with a penalty of extra computation resources needed.

C. Weak Performance Guarantees to In-profile Packets

In the trials, through the use of smaller quota control period and smaller quota amount, the congestion problem in the NTU dormitory network was resolved and the

average packet loss rate in the high priority queue was bounded during a quota control period. However, if any chaotic period took place, the packet loss rate in the chaotic period would be much higher than that in the stable period. To support more consistent performance guarantees, it is essential to control the duration of chaotic period.

We did not consider *credit-carry-over* in the trial (i.e. unused quota not carried over to the next period). The reason is that it would make the traffic demand in each control period even more uncertain and unpredictable. It becomes difficult to preplan and allocate sufficient resources to support performance guarantees to in-profile packets. In Section 6, we use simulations to show the relationship between the important operational parameters of the QPC scheme to provide performance guarantees.

4. Enhanced QPC Schemes

In view of the above-mentioned problems encountered in the field trials using the basic quota-based priority control (QPC) scheme, in this section we propose two different combinations of quota control and priority scheduling methods that achieve better fairness and the support of performance guarantees to in-profile packets using flat-rate unlimited access service plan.

A. Multiple Priority Levels

In a quota control system, the occurrence of chaotic periods is unavoidable. To shorten the duration of a chaotic period, we propose to divide per-user quota allotment into multiple partitions assigned to different priority levels. Each user starts from the highest priority level. As amount of traffic increases, a user is downgraded to one level lower. Depending on the quota amounts assignment to different priority levels, this method can quickly sort users into different usage groups to minimize the performance impact by heavy users to light users.

In *Uniform Quota Assignment* method, assume there are K+1 priority levels (K>1) and the lowest priority level is the best-effort service. The total quota amount Q is equally divided and assigned to the first K priority levels, i.e. each level has Q/K quota allotment. The best-effort service has no quota constraint. All users start service at the highest priority level. As usage accumulates, heavy users will move from the highest priority level to the lowest. The more priority levels, the better in separating different usage groups. As a result, the duration of chaotic periods at each level will be shortened, thus guaranteeing better performance to in-profile packets from light users and fairer sharing of resources. As the number of priority levels increases, the method approximates processor sharing-based scheduling [8].

In *Load-based Quota Assignment* method, we will take into account the distribution of user traffic demands in quota assignment to different priority levels. Consider that the total quota amount is 90 Mbits and there are four priority levels. If a uniform quota assignment is used, each of the first three priority level will have quota allotment of 30Mbits. Suppose regular users only have average 10Mbits usage,

which is much smaller than the allocated quota for the highest priority level, the negative impact from heavy users remains severe. To address the problem, we propose to allocate quota amounts to meet regular users' demand, for instance 10Mbits for level 1, 15Mbits for level 2 and 35Mbits for level 3.

B. QoS Options

By having multiple priority levels, we are able to reduce the duration of and packet loss in chaotic periods by quickly differentiating heavy users from light users. However for users who only occasionally use the Internet still have a non-zero possibility of encountering a chaotic period. To address this issue, we propose best-fit and on-demand QoS-option service models.

1) Best-fit QoS-option Service Model

In the best-fit QoS-option service model, individual user submits an estimate of his/her expected traffic demand to the service provider before each quota control period begins. Consider K+1 service levels. The quality of service of level 1 is better than that of level 2, and so on. Assume non-increasing quota allotments, i.e. $Q_1 \le Q_2 \le ... \le Q_{K+1}$, and non-decreasing charge rates, i.e. $p_1 \ge p_2 \ge ... \ge p_K$. For service level K+1, there is no charge because it is the best-effort service. The virtual Net$ allotment is fixed and the same for each priority level denoted as M, i.e. $p_k \cdot Q_k = \mathrm{M}(\mathrm{Net\$}), \forall k$. Each user i chooses a service level that best matches his/her expected offered load w_i, i.e. $Q_k \le w_i$.

Once determined, a user's account is credited with Net$ M. During a quota control period, when a user's account balance becomes zero, his/her access would be immediately downgraded to the lowest best-effort service.

In this method, users must follow the rule and choose a service level best for their needs to avoid a performance penalty. If a user cheats by giving a smaller than expected offered load and starting with a higher priority level, because the higher priority level has smaller quota allotment, this user will quickly use up the virtual money and be moved to the lowest best-effort queue, possibly experiencing poor performance for the rest of the control period. Directing overloading traffic to the best-effort queue when with empty account can be considered as a penalty to users for possible cheating. On the other hand, it is possible that a user may unintentionally underestimate the demand. This method can encourage users to accurately estimate their usage in order to receive good performance.

The design rationale of this method is to motivate users to better estimate their offered load in each quota control period. Since all users pay the same amount of service fee, if users want to receive better QoS, they must reduce their traffic demands to have higher priority. For users with large-usage demands, they will be served with lower priority to avoid penalty. Under this service model, light users will receive better performance for less use than heavy users. Moreover, the service provider is able to get more information about network load from user selections. This can greatly aid network capacity planning and traffic control. By proper choice

of priority levels and quota allotments, performance guarantees to all levels except the best-effort service is possible. If all users are able to estimate well, the system will achieve good performance and fair resource sharing. In Section 5, simulation results show that users with wrong choice may experience 30% more packet loss.

2) On-demand QoS-option Service Model

In the best-fit service models, users are required to make good estimates about their access demands. However, some users may not be able to accurately forecast their requirements and possible incidental needs for higher quality of service. In [6], the authors reported that in their experiments, although the majority of users chose flat-rate unlimited-access service plans, almost every user purchased high quality service at least once. The objective of the on-demand QoS-option service model is to complement flat-rate service by allowing users to pay an extra fee in order to choose the level of priority service that best meets their QoS requirements. This approach is simple and more predictable than prior researches such as IntServ and optimal pricing policy [3,4]. Accordingly, network service providers may need to allocate extra resources and/or reconfigure the network to accommodate additional usages during a quota control period.

5. Performance Analysis

There are four important parameters in the basic QPC model in achieving fairness and performance guarantees: bottleneck link capacity, quota amount (in bytes), accounting interval (in seconds) and quota control period. In this section, we use Ns-2 [12] simulations to study the choices and relationship between these four parameters. Consider an environment with 100 users: 95 regular users and 5 heavy users. Packet arrivals of all users are Poisson processes with different rates. For the regular user group, the mean arrival rate is 20Kbps with 10Kbps standard deviation. For the heavy user group, the mean arrival rate is 1Mbps with 0.5Mbps standard deviation.

Given a six-hour quota control period and a per-packet accounting interval, Figure 2 compares the amount of successfully transmitted packets with and without a quota control. Users are indexed in the order of their offered loads. Each experiment is run thirty times, and the mean and standard deviation are taken. With a quota control, the throughput performance is significantly enhanced. The throughout line coincides with the offered load curve. For regular users, such performance improvement is owing to the effective reduction of packet loss in chaotic periods. This simulation result confirms the results in the field trials.

In practice, system administrators would like to do usage accounting as less frequently as possible. Figure 3 shows the packet loss rate in the high priority level under different accounting intervals for different quota assignments. The performance guarantees to in-profile packets is to maintain target maximum packet loss rate for the high priority class is 0.01. The packet loss rate performance curves exhibit a staircase shape when increasing the quota allotment. For each curve, the initial increase is due to packet loss occurring in a chaotic period that usually lasts longer for a larger accounting interval, resulting in greater packet losses. In this example, to

achieve the target packet loss rate, the accounting interval cannot be longer than 20 minutes for quota less than 1800Mbits per a 6 hour quota control period

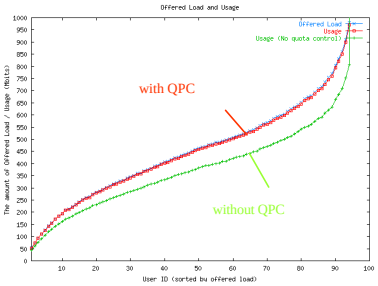


Fig. 2 Comparison of the throughputs with and without QPC scheme

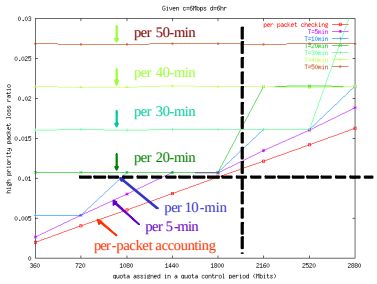


Fig. 3 Packet loss rate vs. quota assignment.

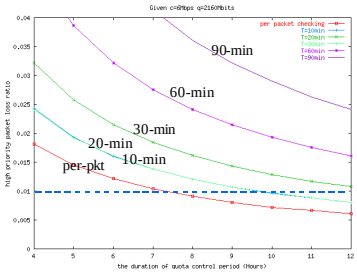


Fig. 4 Average packet loss rate of the high priority queue.

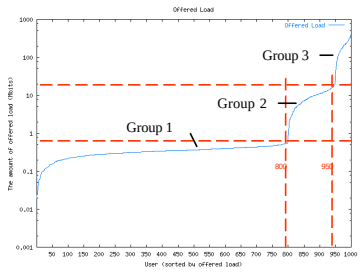


Fig. 5 The user traffic demand distribution.

In this experiment, we are interested to know the performance results under different quota control periods. Figure 4 shows the average packet loss rate for different durations of a quota control period. Because of high packet loss in the chaotic period that occurs at the beginning of a quota control period, a long enough quota control period is needed to average out the packet loss rate. The figure also shows that a shorter accounting interval results in better average packet loss rate.

Table 1 presents the comparison of average packet loss rate of each user group under different number of priority levels. Here, the bottleneck link capacity is 15Mbps, the total quota assignment is 14Mbits, the quota control period duration is 600 seconds and all assuming per-packet accounting interval. Compared with one single service level (i.e. only best-effort service), the packet loss performance of user group 1 is significantly reduced from 20.865% to 0.901% when having two priority levels. This again confirms the results in our field trial. As more levels of control or priority are added to the system, the packet loss rate performance is further improved. As long as a user stays in-profile, a good packet loss performance is guaranteed. For heavy users, packet loss rate becomes very high when entering the best-effort queue.

Table 1. Comparison of average packet loss rate for different user groups.

# of Priority Levels	User Group 1	User Group 2	User Group 3
Best Effort	20.865%	20.810%	19.496%
2	0.901%	1.622%	22.807%
3	0.438%	1.210%	22.879%
4	0.232%	1.062%	22.906%
5	0.135%	0.970%	22.922%
6	0.100%	0.924%	22.929%

Table 2. Performance comparison of two- and three-level of priority.

Quota Assignment Policy: Total Priority Levels	Priority Level	Duration of Chaotic Period (sec)	# of Remaining Users	Packet Loss Rate
Demand-based Quota Assign.: 3 levels	**1**	0.0 ~ 1.7	800	1.1 x10-5
	2	1.1 ~ 44.6	150	2.83x10-3
	3	N/A	50	0.230
Uniform Quota Assign.: 3 levels	**1**	0.0 ~ 14.9	849	4.38 x10-4
	2	11.2 ~ 28.2	82	1.21x10-2
	3	N/A	69	0.229
Uniform Quota Assign.: 2 levels	**1**	0.0 ~ 28.2	931	2.523x10-2
	2	N/A	69	0.228

In the previous experiment, 849 users use less than 7Mbits and remain in priority level 1 after the system becomes stable. Among them, 800 of them use less than 800kbits in total. The variance of usage within a priority level is quite large. To minimize the usage variance with a user group, quota allotment to different priority levels is proposed based on the demand distribution. 0.6Mbits and 19.4Mbps are taken as the quota allotments to priority level 1 and 2, respectively for the demand-based quota assignment. Performance results of each priority level are shown Table 2, compared with the results in two and three priority levels using uniform quota assignment. The demand-based quota assignment approach has the best result in guaranteeing packet loss rate performance to in-profile packets.

In the previous experiments, packet re-classification is based on user account balance. In the following experiments, we consider the Best-fit QoS-option service model with three priority levels. Let q_1=20Mbits, q_2=50Mbits, q_3=infinite and N=1000 users. The traffic demand distribution is shown in Figure 5. The quota control period is 600 seconds. Two experiments are conducted. In the first experiment, all users are assumed to make the right selection. There were 950 users choose service level 1 and the others choose service level 2. The throughput performance of each service level is shown in Figure 6. The x' s at the x-axis denote the time instants at which users are re-classified to lower service levels. In this scenario, the traffic demands of the first 95% users are fully supported and there is no packet loss. The remaining 5% users received 0.2% to 30% packet loss depending on their traffic demands.

In the second experiment, assume a heavy user cheats – the 975th user purposely chooses service level 1 instead of level 2. As shown in Figure 7, before time 63 seconds, the throughput of level 1 is a bit higher than the case of no cheating as shown in Figure 6. At 63 seconds, the cheating user is moved to level 3. Thereafter the performance of level 1 remains stable. For the cheating user, his/her traffic will initially receive good performance but afterwards the performance will be very bad. If the user had not cheated, the average packet loss rate would have been around 22.1% instead of 29.8%. This is the performance penalty for the user. In summary, if all users properly choose their service levels, the overall performance for all users will be good.

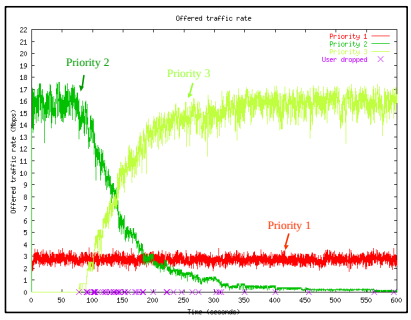


Fig 6 Throughput performance using the Best-fit QoS-option service model assuming all users correctly select service level.

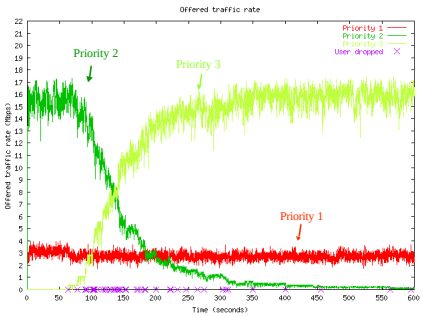


Fig 7 Throughput performance when a heavy user cheats by choosing a higher service level.

6. Conclusion

The main appeal of the flat rate plan is its simplicity as it reduces risk and administrative costs. The unfortunate consequences of flat rate are congestion and no support of performance guarantees. Quality of service cannot be provided to those willing to pay for it. Furthermore, in largely best-effort service based Internet, excessive and abusive use of shared resources by selfish users may severely impair network performance and fairness between users. However, if the consumers regard the pricing structure as too complicated and service providers regard the implementation and administrative costs to be too high, flat rate plans will be preferred.

For the sake of operational simplicity and manageability, instead of using a complex per-flow scheduling approach, in the VIP project we combine a per-user quota control with a priority scheduling scheme. Measurement and performance results from the experiments demonstrate the benefits of this simple scheme. We discuss the problems of chaotic periods, bandwidth stealing, weak support of bandwidth guarantees and the difficulty of using quota incentives to change user network access behavior in quota-based control.

Several methods combining multiple priority levels with quota assignment are proposed to achieve fairer resource sharing, congestion control and support of per-

formance guarantees to in-profile packets. By taking into account user traffic demand distribution in quota assignment to different priority levels, the method can quickly sort users into different usage groups, reducing the duration of chaotic periods and minimizing the performance impact from heavy users on light users. Better fairness and performance are achieved for the popular flat-rate unlimited access plan. Two QoS-option service models are also proposed. In the best-fit model, if all users can properly select the service levels best for their needs, good performance can be guaranteed to all users. This method motivates users to better estimate their offered load in each quota control period. Since all users pay the same amount of service fee, if users want to receive better QoS, they must reduce their traffic demands to have higher priority. For users with large-usage demands, they will be served with lower priority to avoid penalty. In the on-demand QoS-option service model, users can choose their level of priority service to meet their QoS requirements at any time. User can credit the account at any time to receive QoS service charges.

Simulation results of the proposed methods are presented to show the benefits of the simple quota control priority scheduling in resolving fairness, congestion and performance guarantees issues for service networks using flat-rate unlimited access service plan. It successfully provides a more predictive, affirmative service guarantees to the service users.

References

1. R. Cocchi, D. Estrin, S. Shenker and L. Zhang, "Pricing in Computer Networks: Motivation, Formulation, and Example", IEEE/ACM Transactions on Networking, 1993.
2. J. Mackie-Mason, H. Varian, "Pricing Congestible Network Resources", IEEE Journal on Selected Areas in Communications, September 1995
3. S. Shenker, D. D. Clark, D. Estrin, and S. Herzog, "Pricing in Computer Networks: Reshaping the Research Agenda", ACM Computer Communication Review 1996.
4. F. Kelly, "Charging and rate control for elastic traffic," European Transactions on Telecommunications, Volume 8, 1997, pp 33-37.
5. M.D. Biddiscombe, J.E. Midwinter and S. Sabesan, "Application of free-market principles to telecoms resource allocation," Electronics Letters, Vol. 35 No. 4, February 1999.
6. J. Altmann and K. Chu, "A Proposal for a Flexible Service Plan that is Attractive to Users and Internet Service Providers", INFOCOM 2001.
7. T. Lin, Y. Sun, S. Chang, S. Chu, Yi-Ting Chou and Mei-Wen Li, "Priority-Based Internet Access Control for Fairness Improvement and Abuse Reduction," *2nd International workshop on QoS in Multiservice IP Networks* (QoS-IP 2003), Italy, 2003.
8. A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case", IEEE/ACM Transactions on Networking, Vol. 1, No. 3, pp.344-357, June 1993.
9. S. Floyd and V. Jacobson, "Link-sharing and resource management models for packet networks," *IEEE/ACM* Trans. Networking, vol. 3 pp. 365-386, Aug. 1995.
10. S. Blake, et al., "An Architecture for Differentiated Services," IETF RFC2475, December 1998.
11. D. Estrin , and L. Zhang "Design Considerations for Usage Accounting and Feedback in Internetworks", ACM Comp. Commun. Rev., Vol. 20. No. 5. pp. 56-66, October 1990.
12. The Network Simulator, http://www.isi.edu/nsnam/ns/