Fair Usage and Capping for Providing Internet for All in Developing Countries

Yvon Gourhant, Ali Gouta, and Venmani Daniel Philip

Orange Labs, France Telecom R&D, Lannion, France
{yvon.gourhant,ali.gouta,danielphilip.venmani}@orange-ftgroup.com

Abstract. The concept of fair usage is a technique that has existed for years to achieve dynamic network resource allocation when the users do not consume their broadband access continuously all the time. Each user is expected to use his/her Internet access for only a short time or not at full speed all the time. Otherwise they may impair the quality of experience of other users. The purpose of fair usage and capping is to prevent a small range of users from consuming the entire bandwidth allocated by the network operator for all users. In this paper we propose a new fair usage model that aims at satisfying all the actors (OTT providers, network operators, clients on top of the pyramid, mass-market clients). This model is dedicated to developing countries. We implemented it on an open BSD router and measured impact of performances.

Keywords: Developing countries, Fair usage & Capping, Network resources, QoS.

1 Introduction

Fair usage is a technique that was designed and developed to achieve dynamic network resource allocation since research results revealed that the users do not consume their broadband access continuously all the time. Each user is expected to use his/her Internet access for only a short time and/or not at full speed all the time. Otherwise they may impair the quality of experience (QoE) of other users. The idea of fair usage and capping is to prevent a small range of users from consuming the whole bandwidth allocated for all users. Although fair usage was introduced in fixed broadband networks, it is considered to be more relevant with wireless Internet where bandwidth is limited by the scarce radio spectrum. Besides, fair usage is a critical issue for developing countries because it is one important lever of 'Internet For All' which is one of the real challenges for the coming years in the telecom domain. Fair usage is a key issue for building Internet networks based on a design to cost approach. Current Internet offers are mostly targeted to high ARPU (Average Revenue Per User) and the billing is prepaid. This necessitates the need for new solutions comprising both for high and low ARPU customers. The solutions adopted presently in developed countries are not satisfactory because they are based on unlimited offers. This means that the abundance of energy supply, numerous wired access and backhaul resources

R. Popescu-Zeletin et al. (Eds.) AFRICOMM 2011, LNICST 92, pp. 35–48, 2012.

and the larger spectrum in wireless access leads enables them not be bothered about such issues more likely. However, the current solutions let the door open to heaver users that pay the same as users who underuse their Internet access. A simple counter-measure from operators consists to set arbitrarily low priorities to applications that consume lot of resources, such as P2P (peer-to-peer) because there is no way to distinguish low priority traffic more precisely on other criteria (e.g. contents rather than containers).

Therefore, we advocate for solutions that differentiate content distribution and interactive traffic, and that let the user to deal with a given amount of traffic within given periods of time (peak and low hours). We also aim at distinguishing different offers that fit high and low ARPU clients. Thereby in this paper, we present a new fair usage solution for developing countries that we have validated by a proof of concept integrated to an open source router. The paper is structured as follows. The next section gives an overview of economical drivers in developing countries that require enforcing a solution dedicated to these countries. Section 3 presents current fair usage models, research works that may contribute to define new models and the model that we propose. Section 4 describes the implementation of our solution integrated into an open BSD router and shows the impact on performances. Finally, the last section concludes the paper and gives perspectives on following steps to deploy this solution in real networks.

2 Economical Ecosystem and Problem Statement

In order to define appropriate solution for fair usage and capping models for developing countries, it is necessary to have an idea of the local economical conditions for setting up data networks. We noticed that some specificities that show that providing Internet for all is a real challenge. Addressing this challenge is not only the network operators' responsibility but also a global issue concerning energy suppliers and carriers, OTT (Over The Top) service providers, government and regulatory instances, investors, and even end users that should be aware of limited and costly resources. The major cost investments for setting up and maintaining a data network infrastructure may be distributed on the different network partitions. First, there are cases such as Africa where interconnection costs consist currently a large part of total OPEX (Operational Expenditure) costs; they may raise until 50% of the whole OPEX in some countries; nevertheless they are expected to decrease in the next few years by the arrival of new submarine cables. Countries inside Africa have an additional challenge for setting up a PAN-African backbone based on optical fibers. Second, backhaul and long-haul network partitions may also contribute to increase costs in case of long distance links and of lack of existing wired/optical networks because civil costs are expensive; wireless substitution technologies to fiber and copper are microwaves for short and middle distances (based on multi-hop) or satellites for longer distances in remote areas. Therefore, there may be also bandwidth limitations on this network partition. Third, access network costs represent the highest costs due to the number of sites that replicate costs for site renting, civil engineering for setting up base stations towers (in the case of radio networks that is the most frequent), and energy consumption and maintenance. These factors prevent to deploy too much base stations in a given location (an order of idea about radius is 1 to 1.5km in dense areas, 1 to 5kms in urban and suburban areas, 10 to 15kms in rural and larger distances for remote sites). Moreover, the limited frequency spectrum, especially when there are many operators in the same country is a limitation factor for providing more bandwidth. Therefore building a low cost network infrastructure will not be enough. We will need new mechanisms to use these costly resources optimally in a fairly manner. The cost of present Internet offers is afforded only by a small subset of the population representing the top of the pyramid, including small enterprises. These offers are based on networks that are enabled to provide quality of services (QoS). There is a need to penetrate the Internet access to the rest of the population. This gives an opportunity for telecom operators to reuse their network infrastructures for providing best-effort Internet offers if new offers do not jeopardize existing offers.

3 Fair Usage and Capping Model

3.1 Current Situation in Developed and Developing Countries

In developed countries such as the western European countries, Japan, the U.SA, etc., data mobile rates are falling into unlimited offers drawn by flat rate models in xDSL networks. The inability to create a standard business model that would make every actor to be satisfied is the current situation in such countries. OTT players claim network neutrality whereas network operators declare that network resources consumed by OTT services cost more than final users pay for their traffic consumption. On the user side, we could notice that the resources consumption can be depicted in a Pareto law where 10% of clients are consuming nearly 80% of network resources (see Fig. 1). The users who consume a large amount of the bandwidth claim that they just consume network resources they are allowed to. Some network operators arbitrarily reduce the priority of some applications, such as P2P applications seen as background applications but sometimes erroneously since there is no real standard classification of these applications. Without knowing the end-user point of view while giving such priorities, the impact for the user is not easy to be determined. Therefore, distributing the network resources among clients based on application-level bandwidth management is not well perceived by the end-users. It raises a competition between application-signature recognition in the network by deep packet inspection (DPI) equipments on one hand, and on the other hand, camping new techniques for masking application identity (http masquerading, encryption, dynamic behaviors, changing protocol TCP to UDP, changing ports, etc.).

In developing countries, most offers are based on prepaid, and quotas are set-up to limit the network resource consumption. When a user reaches a quota associated to his/her account then a rate limit is applied or all the associated data traffic is blocked. This is a too simple scheme since the user may be frustrated to

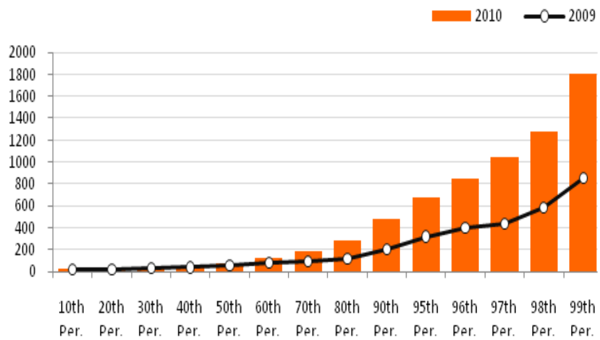

Fig. 1. Distribution of volume among users (sources Cisco)

overtake quota even when the network is not fully loaded. Moreover, quota should not be a single solution because there are some traffic that may be counted in quotas even though they should not be included (e.g., end-user device firmware update, end-to-end retransmission, . . .).

Therefore, this paper aims at presenting new fair usage mechanisms that would make every individual actor satisfied. We assume that the end users are totally aware that the network resources that they are provided with, comes with a cost, as this is the case in developing countries.

3.2 Related Works

In countries like Tunisia, India, etc. quotas are limited according to the periods of time. Network operators distinguish quotas/rates at day time and at night (e.g., unlimited offers limited at night). This is the first step towards the direction that would benefit all but our opinion is that it is not enough. Heavy applications may still grab network resources at peak time until quota thresholds are reached. Applying yield management techniques for providing discount at appropriate time and locations when a cell is not loaded could also be considered as a possible solution. These techniques have succeeded for voice traffic (e.g. traffic-zone, bonus-zone which broadcasts cell-by-cell discounts periodically by Unstructured Supplementary Service Data (USSD) according to the network load of the cell). But these techniques need to be revisited for data which is more complicated. Because, in data networks, network and service may not be provided by the same actor, the network operator does not know service priorities. Different applications have different bandwidth constraints in terms of throughput, delay and packet loss. We consider yield management as a possible solution but we postponed it since it needs first to set up a simpler fair usage and capping model.

Some of the recent works prove that there are already published results on the benefits of using age based scheduling policies on networks to improve user perceived performance [1,2,3,4,5,19]. Priority based scheduling falls in the general framework proposed by Ruschitzka and Fabry [6]. Processor Scheduling (PS) policy has been studied in [7] and compared to round robin. PS is extended to Generalized Processor Sharing (GPS) [8] and Discriminated Processor Sharing (DPS) [9] to support weighted sharing. The LAS/FBPS/FB policy is first studied by Schrage [10] but has received significant recent attention for the case where the task sizes have a large coefficient of variation [11]. Several other blind scheduling policies, such as multi-level feedback-queue scheduling (MLFS) [12], Multi-level Processor-Sharing (MLPS) [13] and its special cases [14] have been also investigated in the context of computers. The scheduling mechanism RuN2C (Running Number 2 Class differentiation mechanism) in [15] gives priority to connections that have small number of packets without penalizing long flows. Its authors suggest setting a threshold that should be well tuned and give two classes of priorities: high and low priority. Scheduling mechanisms defined previously in order to find a compromise between long and short flows suffered from a huge number of parameters that need to be tuned. RuN2C proposes a lighter version in which there is only one threshold to be parameterized and thus distinguish between long and small flows. RuN2C associates the goods of LAS and PS. We had a particular look at [16, 17] because the two different types of age based scheduling policies, LAS and MLPS, are using similar criteria that we looked for. The concept of the Least Attained Service (LAS) scheduling policy is based on size files scheduling. But there is no prior knowledge of a job size, and packets are forwarded according to the amount of processing time that was given to that job. The Multi Level Processor Sharing (MLPS) scheduling policy consists on setting several thresholds; when the number of packets belonging to one job exceeds one threshold this job sees its priority being decreased. So the biggest the size is the more the priority decreases. Finally, authors in [18] quantify benefits and drawbacks related to the deployment of per-flow scheduling (Fair Queuing, Longest Queue First, Shortest Queue First) related to TCP and UDP protocols. Shortest Queue First has an attractive property to implicitly differentiate streaming and interactive traffic, performing as a priority scheduler for applications with low loss rate and delay constraints. In such a way it achieves a similar objective than our solution at a lower cost but further works need to be done to know the impact on heavy flows since it applies to burst periods only. All these works focus on scheduling whereas requirements from the field in developing countries require adapting policies during the day. Our model tries to tackle this.

3.3 Proposed Model

Our proposed solution consists to apply different fair usage policies at given periods of time. We noticed that fair usage is particularly crucial at peak hours because a large proportion of the network resources are dedicated for a short period of time as in fig. 2. If users agree to postpone heavy traffic, especially background applications, after peak periods, then network dimensioning may decrease drastically. The impact on the reduction of the network costs may be seen on the different network partitions, investments for more capacities may be

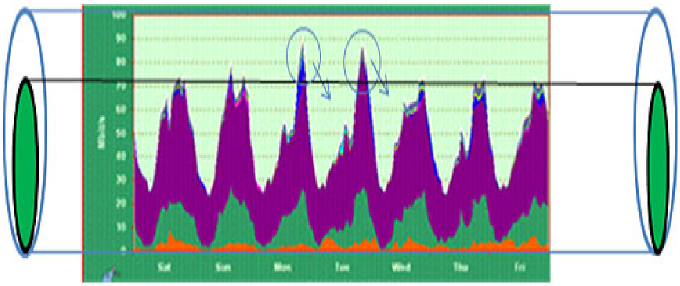

Fig. 2. Evolution of network resource usage during a week (i.e. ratio of traffic load of the aggregation network in y-axis as time is getting on from Saturday to Friday in x-axis). Colors represent different kinds of applications but are not relevant in this context; all applications are cumulated. If some flows (e.g. background applications) can be postponed after peaks, then network operators may reduce Internet prices because they can postpone next investments needed by traffic increase.

postponed by several months. This is really different to the traditional scheduling mechanisms such as shaping which delays some packets in few milliseconds and therefore it appears to be focusing more on dealing with traffic bursts.

One major difficulty is to find a way to set priorities between applications that satisfy the users and the network operator. Although there exist user-network signaling protocols such as Resource Reservation Protocol (RSVP), these protocols are not implemented by Internet applications and not activated in networks for this purpose. So, we can't rely on such a protocol to define user priorities.

Our aims are twofold. On one hand, we suggest to give a better priority to interactive applications during peak hours, and to give better conditions to streaming and download applications at low hours. On the other hand, we consider different user profiles, such as gold/silver/bronze, in order to monetize bandwidth, giving more bandwidth and better rate limit to gold users who pay more. Our solution is based on two criteria: max rate and quotas. Gold users will be granted of a better max rate and a bigger quota than best effort users. Fig. 3 depicts the global schema for our proposed fair usage mechanism.

We allocate a quota and a max throughput rate per user and per period of time taking into account his/her profile. We define two periods of time (peak and low hours) and different bandwidth management policies for each of them. The dissatisfaction that is caused at peak hours will be compensated at low hours, typically upgrading best effort users to gold users at low hours. The impact will consist for instance to reduce the download of a file at a peak hour and to increase it at low hours.

At peak hours, (1) the scheduling policy gives better priority for smaller flows; the priority of more voluminous flows will decrease in two times, respectively by putting them into a best-effort queue, then into a less-than-best-effort queue; we will see in the next section how we pick up these flows and how we have fixed threshold values. (2) We also decrease the rate limit as far as the quota is decreasing for best effort users. There is no impact on gold users clients. Moreover, our model gets rid of the need and complexity to display the fare period. Reciprocally, our model includes incentive mechanisms for applications that use protocols which stop transmission whenever they encounter congestion (instead of retransmitting), such as Lower-than-Best-Effort Transport Protocols (LEDBAT). We have included special compensating mechanisms for these applications as well by granting extra-quota at low hours.

At low hours we differentiate gold and best effort users by giving them different rate limits. At the beginning at the low hours, we increase quotas for users that have been quiet at peak time, and we remind the best-effort users that have been impacted at peak hours in order to grant them the same quota than gold users. During low hours, if a quota expires then the user rate limit value is decreased. If order to reduce the impact on interactive applications, there is need to apply a scheduling policy such as SQF (Shortest-Queue-First) that has been discussed in the subsection 3.2.

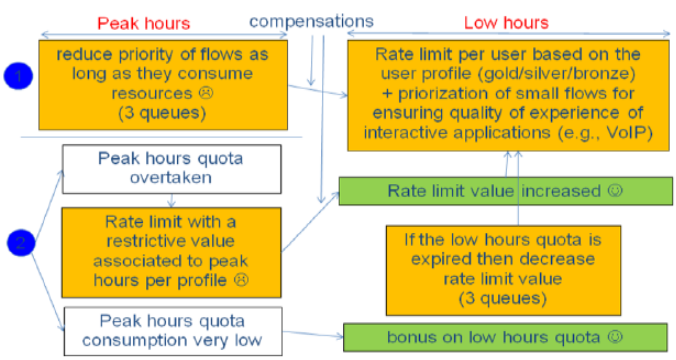

Fig. 3. Schematic diagram for fair usage considering peak and low hours

4 Implementation Issues

4.1 Feasibility Study

As a proof of feasibility, we implemented the schema described above on a real router (OpenBSD). At first we focused on reducing the priority of flows that exceeded a certain volume; the threshold has been set up by trail and error (see subsection 4.3). Then in a second attempt we applied a rate-limit to the traffic of users that exceeded their quotas.

4.2 Reducing Priority of Flows

We classify the flows based on their volume with respect to the consumption of bandwidth at IP level (no need for application-level analysis). We defined a set of queues with different weights (time service) and then according to the volume of a flow through the time we decide to which queue this flow will be forwarded. We maintain a state on the flows that goes through the router. For each incoming packet we look for the state on the flow that is maintained so that we can update the corresponding values (number of packets, volume etc.). We made this choice because OpenBSD uses a strong stateful firewall named Packet Filter (PF). In order to maintain states on connections PF uses Red and Black trees that are typically in-memory structures used to provide fast access to the memory where all states are stored. In this case, R&B trees ensure a fast lookup (if a new packet goes through an interface), insertion (if a new packet of a new connection goes through an interface) and deletion (if a connection is supposed finished, i.e. the packet has got a FIN flag in case of TCP connection). All these operations are about O(log(n)) and this is due to the R&B tree implementation. PF can work in a stateless mode as well as in stateful mode, but its design works in stateful mode by default. We defined filtering rules in the configuration file of the firewall: when a packet matches a rule, then a state is created; all the following packets of the same flow will not check the set of rules; however, they will check the R&B tree to look for the corresponding state entry. Fig. 4. shows these states on flows.

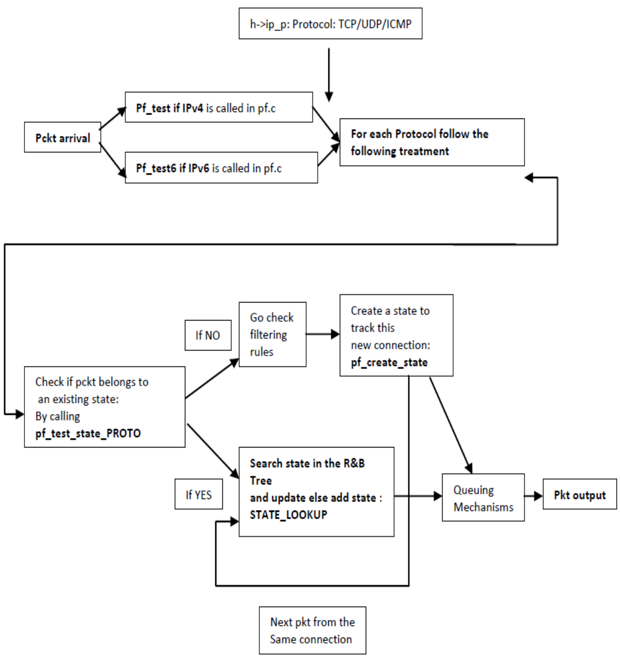

Fig. 4. Description of Packet Filter maintaining the states on flows

OpenBSD uses ALTQ (Alternate Queuing) mechanisms to provide queuing disciplines and other QoS related components required to achieve resource sharing between queues. In the configuration file of the firewall, we defined queues and the scheduling mechanism. Fig. 5. represents the flow diagram for tagging

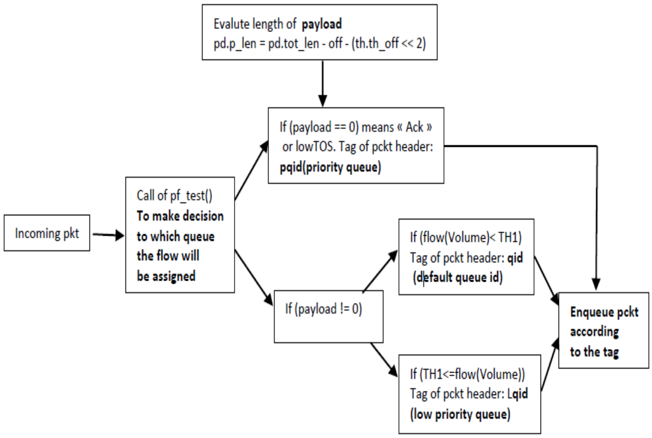

Fig. 5. Flow diagram for tagging packets with the ID of a queue to distinguish the long flow from the small flow according to the threshold TH

packets with the ID of a queue in order to distinguish the long flows from the small flows according to the threshold TH.

We defined queues having different priorities and we queue packets according to the volume of the flow in such a way that we can promote small flows. The issue is that there is no way to re-assign a flow to a different queue. It always gets assigned to one queue which is, the first packet of a flow. So we wrote a process in the firewall that achieves the following behavior: when a new packet comes, after looking for the corresponding state in the R&B tree, the volume of this flow is computed and a tag in the header of that flow is set with the ID of the lower priority queue. All the following packets belonging to that flow will get the same tag. In the next step, we shape the traffic of users that exceeds their quota. In order to avoid to implement a full subscriber management system, we distinguish user profiles (Gold, Best Effort) by referring to their IP addresses in order to simplify the proof of concept. Even addresses represent the Gold profile and odd addresses represent Best Effort profile. The differences between the two profiles are the value of the quota threshold and the value of the rate limit. We implemented two schemes (algorithms)(fig6, fig7) to change the rate-limit associate to IP addresses.

We count the traffic of each IP address and when we find that one IP address exceeded its quota, we add this IP address in a table of exceeders. When a new packet comes, we first check this table to see whether the IP address is listed in this table or not. If it exists then we assign this packet to a Queue which we give the lowest serving time. If it doesn't exist we then check the filtering rules to determine to which queue this packet will be assigned. The following scheme is an example of shaping: assigning traffic according to the profile of the user (Gold or Best Effort), the shaper has 15% of the available bandwidth 10% for Gold profile and 5% for Best Effort profile.

Fig. 6. Create one queue for Gold profile and one queue for Best Effort profile

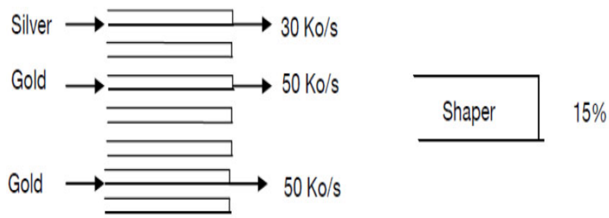

Fig. 7. Create a new queue per IP address and associate its rate limit to the user profile

We make the usual count of traffic per IP address. Then if we detect that a user has exceeded his/her quota, we create a new queue in the configuration file of the firewall and we assign all eventual connections triggered by this IP address to its "own" queue. Then we apply rate limit to the outgoing traffic to a fixed throughput value associate to the user profile. Typically, OpenBSD provides Hierarchical Fair Curve Service (hscf) in order to limit the throughput of a queue with a upper-limit value (e.g. max attained throughput) which the queue will not exceed even if there is a situation when more bandwidth is available. Therefore, if a customer exceeds his/her quota, we first create a new queue on the fly that uses that upper-limit feature of hfsc, then all packets that come from this IP address they will be assigned to that queue. To distinguish the Gold profile from the Best Effort profile we adjust the upper-limit value, respectively 30 and 50 kbytes/s in our proof of concept. Real values will depend on the resources of the operational network dimensioning and the operator strategy. The second scheme is more attractive, but the problem is that the number of filtering rules will become so huge at large scale and this may overload the firewall.

4.3 Performance Issues

In order to evaluate our fair usage model and this implementation, we have captured all the traffic between a DSLAM and a BRAS between 3pm and 4.35pm on a Tuesday on the basis of 2000 clients connected. Then we replayed this captured traffic through the interfaces of the router using 'tcpreplay'.

We dumped the outgoing traffic from the interfaces of the router and compared it with the initial capture using the 'tstat' tool which gives statistics about each flow.

We first used the 'tstat' tool to have a primary idea about the volume, duration and number of connections at that time. We looked at the distribution of flows according to their volume and looked for the duration of flows according to their volume so that we can have an idea about the length of bulky flows and analyze in the mean time the Internet traffic.

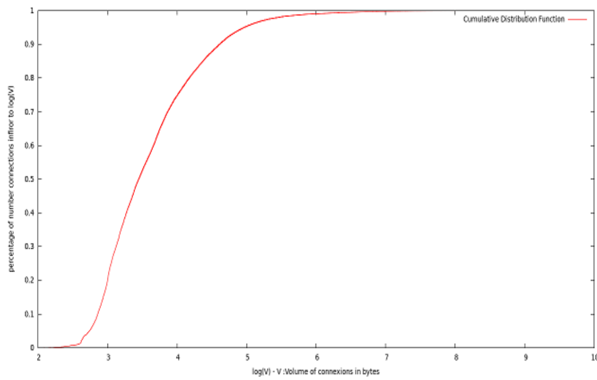

Fig. 8. Repartition of connections according to their volume

Fig. 8. shows the distribution of flows according to their volume within an hour and forty minutes of traffic. Statistics give that there were around 2 millions connections during this duration. We can notice that only 5% of all connections are above 100KB. This is due to basically to the http bursty behavior, e.g. when a client requests a server to open a page. During this time, a lot of connections would be opened just to download some html code. In our case we give interest mostly to bulky and the results underline the assumptions we made for setting up our fair usage scheme.

Fig. 9 shows the duration of flows according to their volume so that we can have an idea about the flows that requires more resources. It is very clear from fig. 7 that the 5% of flows we discussed above are mainly the flows that require most of the resources. Most of them ends in a brief time although they are supposed to be bulky flows.

Then the final test evaluate the impact on performance of reducing the priority of bulky flows so that they either will be dismissed or will be delayed, in order to give more priority to smaller flows. To make this, we replayed the capture of traffic through our OpenBSD router on which we implemented the schemes described in the 4.2. Hence, we observed that the router succeeded to handle more than 10000 connections at the same time without any major decrease in the performance of the router (Fig 10 and Table 1). We also succeeded to set the priority of flows that exceeded our threshold limit and thus redirecting them to less served queues.

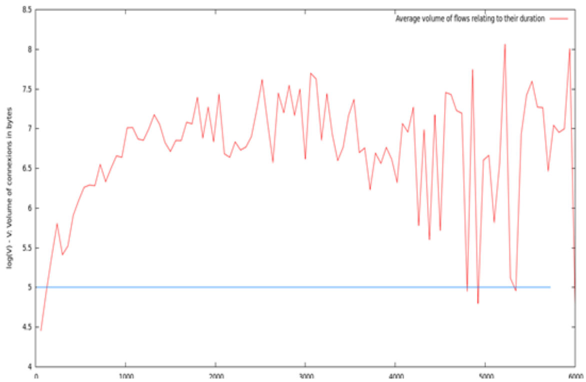

Fig. 9. Distribution of the length of flows according to their volume

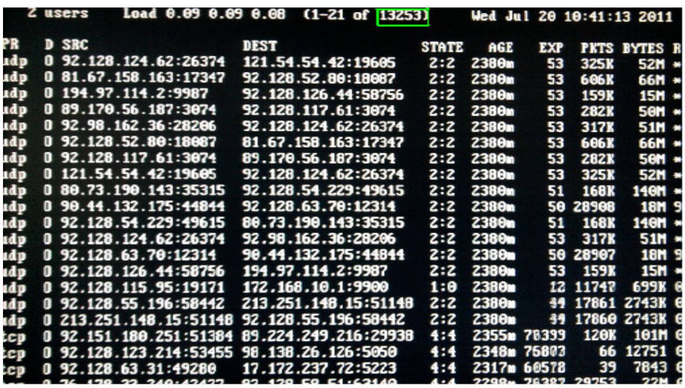

Fig. 10. Capture of the router console showing the impact of the number of connections (in green) on load

Table 1. Impact of changing tcpreplay rate on the router load

Average rate	Number of simultaneous connections	Load(on average every minute)
130Mbps	40000	15%-11%
257Mbps	56000	15%-11%
330Mbps	70000	18%-11%
491Mbps	11000	20%-14%

5 Conclusion

Free access and unrestricted demand for a finite resource ultimately dooms the resource through over-exploitation. This occurs because the benefits of exploitation accrue to individuals or groups, each of whom is motivated to maximize use of the resource to the point in which they become reliant on it, while the costs of the exploitation are borne by all those to whom the resource is available. Before one starts criticizing operators/ISPs for the 'unfair' policy, one needs to understand that bringing a cap to subscriber number per means a direct hit in their revenue and if they need to maintain QoS, they also need to ensure that subscribers too maintain the quality of usage. Through our results, we have proved that it still possible to offer atleast the minimum bandwidth for all the users for certain "given" period of a day that would make every individual actor on the Internet satisfied. However, imposing contention ratio does not help to solve 100% of the problem, but it surely is first step in the right direction. What ISPs need to also ensure is that broadband users are being educated on the usage and repercussions of over-usage and most importantly, quantifying what is fair-usage.

References

1. Massoulie, L., Roberts, J.: Bandwidth sharing: objectives and algorithms. In: INFOCOM, vol. 3, pp. 1395–1403 (1999)
2. Guo, L., Matta, I.: Scheduling flows with unknown sizes: Approximate analysis. In: Proceedings of ACM SIGMETRICS (Extended Abstract), pp. 276–277 (2002), extended version available as a Boston University Technical Report BU-CS-2002-009
3. Bonald, T., Proutiere, A.: Insensitive bandwidth sharing in data networks. Queueing Systems 44(1), 69–100 (2003)
4. Yang, S.C., Veciana, G.D.: Enhancing both network and user performance for networks supporting best effort traffic. TON 12, 349–360 (2004)
5. Avrachenkov, K., Ayesta, U., Brown, P., Nyberg, E.: Differentiation between short and long TCP flows: Predictability of the response time. In: Proceedings of INFOCOM (2004)
6. Ruschitzka, M., Fabry, R.: A unified approach to scheduling. Commun. of the ACM 20(7), 469–477 (1977)
7. Coffman, E., Muntz, R., Trotter, H.: Waiting time distribution for processor-sharing systems. Journal of the ACM 17(1), 123–130 (1970)
8. Parekh, A.K., Gallager, R.G.: A generalized processor sharing approach to flow control in integrated services networks: The single-node case. IEEE/ACM Transactions on Networking 1(3), 344–357 (1993)
9. Fayolle, G., Mitrani, L.: Iasnogorodski. R.: Sharing a processor among many job classes. Journal of the ACM 27(3), 519–532 (1980)
10. Schrage, L.: The queue M/G/1 with feedback to lower priority queues. Management Science 13(7), 466–474 (1967)
11. Rai, L.A., Urvoy-Keller, G., Vernon, M.K., Biersack, E.W.: Performance analysis of LAS-based scheduling disciplines in a packet switched network. In: Proc. ACM SIGMETRICS 2004, pp. 106–117 (2004)
12. Silberschatz, A., Galvin, P.B., Gagne, G.: Applied Operating Systems Concepts. John Wiley & Sons (2000)
13. Kleinrock, L.: Queueing Systems, vol. I: Theory, vol. II: Computer Applications. John Wiley&Sons (1975/1976)

14. Aalto, S., Ayesta, U., Nyberg-Oksanen, E.: Two-level processor-sharing scheduling disciplines: Mean delay analysis. In: Proc. ACM SIGMETRICS 2004, pp. 97–105 (2004)
15. Brown, P.: Stability of Networks with Age-Based Scheduling. In: INFOCOM 2007, pp. 901–909 (2007)
16. Ayesta, U., Brown, P., Avratchenkov, K.: Differentiation between Short and Long TCP Flows: Predictability of the Response Time. In: INFOCOM 2004 (2004)
17. Altman, E., Barakat, C., Laborde, E., Brown, P., Collange, D.: Fairness Analysis of TCP/IP. In: Proceedings of IEEE Conference on Decision and Control, Sydney, Australia (December 2000)
18. Carofiglio, G., Muscariello, L.: On the impact of TCP and per-flow scheduling on internet performance. In: Proceeding of 29th Conference on Information Communications, INFOCOM 2010 (2010)
19. Yang, C.-W., Wierman, A., Shakkottai, S., Harchol-Balter, M.: Tail asymptotics for policies favoring short jobs in a many-flows regime. In: SIGMETRICS/Performance, pp. 97–108 (2006)