Home is Where the (Fast) Internet is: Flat-rate Compatible Incentives for Reducing Peak Load

Parminder Chhabra, Nikolaos Laoutaris and Pablo Rodriguez
Telefonica Research
Barcelona, Spain
{pchhabra, nikos, pablorr}@tid.es

Ravi Sundaram
Northeastern University
Boston, MA, USA
koods@ccs.neu.edu

ABSTRACT

Internet traffic to homes is surging, driven by the demand for rich content and the proliferation of home networks. This creates a huge problem for ISPs since residential customers expect the certainty of a fixed bill while ISPs do not want to upgrade backhaul equipment frequently in the absence of extra revenue streams. We consider simple variants on existing flat-rate schemes that will enable homes to self-select a portion of their peak hours traffic and move it to non-peak hours to benefit from offered incentives. We present a well-defined formulation of the problem and characterize its computational complexity. We show that a simple fractional algorithm achieves the optimal traffic reallocation and is realizable with small modifications to existing infrastructure. The fractional model also captures the reality that homes may be willing to move a fraction of their delay-tolerant traffic in response to appropriate incentives. Using trace-driven simulations based on well-accepted utility models and actual backbone traffic from a large ISP, we demonstrate that our incentive scheme can substantially lower peak congestion while still satisfying the increased demand of home networks.

Categories and Subject Descriptors: J.M [Computer Applications]: MISCELLANEOUS, D.4.4 [Communications Management]: Network communication

General Terms: Algorithms, Experimentation, Measurement.

Keywords: Incentives, Delay tolerance, Flat-rate compatible, User greediness.

1. INTRODUCTION

Motivation. A typical home network connected to a broadband ADSL line supports a variety of applications, each with different characteristics and requirements - email, web browsing, P2P traffic, VOIP, video streaming, gaming, etc. A recent study of packet traces in a home [9] showed that household bandwidth resources are already insufficient and highlighted the tension between different applications (e.g., streaming/gaming application suffering in the presence of P2P). The problem is worse in homes experiencing growth in demand for bandwidth for a variety of reasons - children of the Internet age, availability of HD content for films and TV shows, home computing devices such as telepresence, smart-metering, mobile and location-aware devices etc. As a result, ISPs facing a huge peak-time crunch are forced to respond with stifling counter-measures (e.g., throttling P2P traffic [12]).

In a complex and evolving world, a natural question to ask is whether there exist general principles for alleviating the situation? A substantial body of work at the intersection of computer science and economics has established that incentives are a powerful way to allocate scarce resources. Sophisticated usage-based pricing schemes [3] present strong arguments in favor of managing a network via incentive schemes.

Flat-rate Compatible Incentive Schemes. Unlike businesses that may experience revenues in line with traffic growth, homes have a strong inbuilt preference for stability in their consumption costs. Hence, flat-rate pricing is the de-facto standard for retail bandwidth - a scheme that lacks any incentive for households to make rational usage decisions, especially during peak hours.

Our incentive scheme has its genesis in our position paper [10] where we employed algorithmic and measurement based analysis to make the case both for an incentive scheme as well as a storage augmented scheme to alleviate peak hour from delay tolerant bulk traffic. We develop on the incentive theme here. Our scheme rewards households that self-limit their consumption during peak hours, by granting them higher-than-purchased access rates during non-peak hours, when network is under-utilized due to commonly observed diurnal patterns in aggregate traffic load. The awarded higher-than-purchased rate is designed to (far) out-weigh the rate decrease during peak hours, and thus facilitates a much higher daily download volume than permitted by the purchased nominal rate stated in the end-user's contract. By explicitly soliciting the voluntary participation of homes such a scheme avoids the current situation where (a) users fear that they are being subject to undesirable traffic engineering, (b) users don't make rational use of network resources and (c) ISPs invest in expensive traffic identification and shaping equipment.

Delay Tolerance. The key to our approach is the recognition that a lot of traffic on the Internet is delay tolerant. As an example, consider P2P and One Click Hosting (OCH [2]) downloads of high resolution movies. Such downloads take substantial time and consume valuable resources during peak hours. We analyzed data from a very popu-

HomeNets 2010, September 3, 2010, New Delhi, India.

lar torrent [8] which, although not video material, gave us a sense for existing delay tolerance. This trace clearly validates the widely held belief that users remain seeders for several hours after a download is complete. An analysis of this trace showed that the users remained seeders in excess of *six and a half hours* after the completion of the download. Such idle time is attributed to users leaving the torrent, not at the time of download completion but, at a later time when they are able to access their computer (to presumably consume the content) and thus gives an idea of their delay tolerance.

Neflix offers Netflix Queue to the Xbox service [1] that allows movies to be delivered to the Xbox rather than via snail mail. This means that users are willing to tolerate many hours of download before watching the content at their conveneince.

1.1 Our Contributions

By allowing a home to voluntarily defer peak consumption in return for an incentivized consumption during non-peak hours both the home and the ISP can benefit by deferring infrastructure upgrades. Observe that from a business stand-point the unit of a home allows for the effective use of an in-kind incentive as a tool of behavior change. A single application/user may not be able/willing to consume the incentive, but an aggregate of a household will be able to more effectively consume the incentive via coordination among the household members [4].

We provide comprehensive models for characterizing users and ISPs and study the complexity of finding the optimal incentives. We argue that the most relevant case is one in which ISPs can offer customized bids and households can respond by selectively moving a portion of their traffic and present a highly practical algorithm. We analyze several days worth of aggregate traffic from two customer-provider links of a large transit ISP and focus on the relationship between consecutive load peaks and valleys. Based on these relationships, we observe that even in links with substantial amount of base traffic across all hours, there remains substantial room for expansion ($\sim$2 or more) that can be used for implementing the proposed incentive scheme.

Our contributions are as follows (road-map):

- a precise formulation of the model, the problem and characterization of its computational complexity (Section 2).
- two models for capturing the greediness of users (Section 4.4)
- an empirical evaluation of maximum expansion allowable in aggregate (Section 5.1)
- empirical evaluation of minimum expansion using two different models for disaggregating traffic - uniform and Pareto (Section 5.2).

A small fraction of users may be unable to enjoy the reward of expanded bandwidth due to technical reasons such as last mile access bottlenecks just as there will be a few users who will not accept any amount of incentive. Our assumption is that such users are captured by our greediness models. The analysis validates our fundamental thesis that flat-rate compatible incentive schemes constitute an easy-to-adopt mechanism enabling more efficient utilization for users.

2. MODEL, BIDDING COMPLEXITY

First we develop the problem and its notation in its full generality. Subsequently, we show that the general models are computationally infeasible and restrict ourselves to a simplified model.

Slots and Traffic: Consider a link of capacity C carrying traffic from n homes. We assume that a day is divided into S slots and traffic from an individual home is capped at E.

We divide a day into two intervals: the *peak hour*[1] of the link which is the time interval during which the aggregate traffic load L pushes the utilization L/C above a threshold θ; and the rest which constitute the non-peak hour.

User: Given coordination among household members [4], we assume a user to be a unit of a home. In the general slotted case we will denote the amount of traffic generated by the user $i, 1 \leq i \leq n$ by a vector $\vec{E_i}$, with $E_i(j)$ denoting the traffic generated by that user in the jth slot, $1 \leq j \leq S$. We assume that the user will be willing to accept any one of a constellation $C_i = \{\vec{E_i^*}\}$ of vectors as a substitute for its current traffic pattern, $\vec{E_i}$, given appropriate incentives. In general each vector in the constellation will be an expansion of the original traffic and will represent the incentive demanded by the user in return for the move. In the simplified model we will abuse this same notation slightly and use E_i to denote the amount of elastic traffic generated by the ith user during peak hour, $1 \leq i \leq n$. We assume that a user can move her elastic traffic to the subsequent valley given adequate incentives. To move elastic traffic E_i away from the peak hour, the user demands that she be allowed to send $E_i^* = w_i \cdot E_i$ during the non-peak hours, where w_i is the *expansion ratio*. In the simplified model, we consider two models for the amount of elastic traffic that users are willing to move: in the *fractional* model, users can move any fraction of their elastic traffic; in the *all-or-none* model users move all of their elastic traffic E_i or none at all.

ISP: For the ISP we consider two models: in the *omniscient* model the ISP knows the w_i's of the users whereas in the *oblivious* model it does not.

Problem formulation: Given S slots, a peak traffic bound of E and n users with expansion ratios w_i and traffic vectors $\vec{E_i}$, $1 \leq i \leq n$, is it possible to reschedule users' traffic by satisfying their expansion ratio requirements while maintaining a maximum congestion of at most E?

There are potentially two kinds of ISPs to consider. An omniscient ISP can make customized offers (bids) to each individual client i as an incentive to move elastic traffic to non-peak hour. An oblivious ISP has to select a single expansion ratio to be offered to all users. The oblivious ISP is highly inefficient since it cannot price-discriminate to extract the greatest utility from the system.

The omniscient ISP is the more natural model since ISPs (like every other vendor) offer different deals to different segments of their customer base. We first consider the all-or-none user and show that the problem is computationally tractable only in a very restricted case and even then it is not practically feasible. From a practical standpoint too, the fractional user case is the more relevant one - remember that we are talking of moving delay-tolerant traffic like movie downloads etc and since the user is tolerant of delay it must be acceptable to download a fraction of the content during peak hour and the remainder during non-peak hour. Therefore we focus our attention on the omniscient

[1]The use of the term "hour" may be misleading. In this paper we use it to denote a contiguous sub-interval of the day that may be more or less than 60 chronological minutes.

Slots(S) \ Traffic(E)	Unbounded	Bounded
Unbounded	Strongly NPC	Strongly NPC
Bounded	Weakly NPC	Polynomial

Table 1: Optimal all-or-none bidding.

fractional case and present a simple greedy algorithm that computes the optimally efficient reordering of traffic.

2.1 All-or-none User

We consider four different models depending on whether the number of slots S and the traffic cap E is *bounded* (i.e., fixed independent of n the number of users) or *unbounded* (i.e., grows as a function of n). Table 1 summarizes our results. The strong and weak NP-completeness results follow by reduction from Bin-Packing and Partition, respectively. Due to space constraints we defer the proofs of the entries in Table 1 to a full version of this paper. We state the theorems without proofs below:

THEOREM 2.1. *If S the number of slots is unbounded (i.e., grow as a function of n) then the bidding problem is strongly NP-complete.*

THEOREM 2.2. *If S the number of slots is bounded (i.e., a fixed constant independent of n) and E can grow unbounded then the bidding problem is weakly NP-complete.*

THEOREM 2.3. *If both S and E are bounded (i.e. fixed constants independent of n) then the bidding problem is solvable in polynomial time.*

The bottom-line is that the only case when the problem is tractable is when both S and E are bounded, but even in this sub-case the solution requires complex algorithms involving fixed-dimension integer programming that render it practically infeasible.

2.2 Fractional user

In this section, we discuss the case of the omniscient ISP with a fractional model of the user. Let E_i^o denote the ISP's offer (or bid) to user i and E_i^m the amount of elastic traffic moved to non-peak hours as a consequence of this bid. In the all-or-none user model the user will move $E_i^m = E_i$ iff $E_i^o \geq E_i^*$. In the fractional model the user will move $E_i^m = E_i \cdot \min(1, E_i^o/E_i^*)$ for any bid $E_i^* > 0$.

The ISP's objective is to make bids that keep the peak hour load (utilization) below θ and fit in the valley of capacity V, while minimizing the extra bandwidth spent in incentivizing users to shift their elastic traffic. This objective is justified by the fact that the ISP can use any leftover non-peak bandwidth to support other non end-user traffic, e.g., bulk transfer of scientific datasets and corporate backups [10]. The ISP's objective can be formalized as follows:

Select E_i^o to minimize the excess rate:

$$\sum_i (E_i^o - E_i^m) \cdot I_{\{E_i^m > 0\}} \tag{1}$$

Subject to the constraints:

$$L - \sum_i E_i^m \leq \theta \cdot C \tag{2}$$

$$\sum_i E_i^0 \leq V \tag{3}$$

We characterize the computational complexity of the problem by the following proof sketch.

Algorithm 1 Optimal fractional bidding

function $\{E_i^o\}$=BIDDING$(\{E_i\}, \{E_i^*\}, \theta, C, L, V)$
    $\{E_i^o\} \leftarrow 0_{|\{E_i^*\}|}$
    while $L > \theta \cdot C \quad \& \quad V > 0$ **do**
        $j \leftarrow \arg\min_i \left(\frac{E_i^*}{E_i}\right)$
        $E_j^o \leftarrow E_j^* \cdot \min(1, \frac{L - C\cdot\theta}{E_j})$ $\triangleright$ ISP's bid to user j
        $L \leftarrow L - E_j \cdot \frac{E_j^o}{E_j^*}$ $\triangleright$ elastic traffic moved
        $V \leftarrow V - E_j^o$ $\triangleright$ the valley filling up
        $\{E_i\} \leftarrow \{E_i\} \backslash E_j$ $\triangleright$ do not consider j again
        $\{E_i^*\} \leftarrow \{E_i\} \backslash E_j^*$
    end while
    return $\{E_i^o\}$
end function

THEOREM 2.4. *The bidding problem where the ISP is omniscient and users are open to moving a fraction of their traffic is solvable in polynomial-time.*

Proof: The omniscient fractional model can be solved in polynomial time employing a greedy algorithm. The ISP sorts the users in increasing order of their expansion ratio w_i and bids according to their demands E_i^* until it satisfies the QoS constraint of Eq. (2). The details are in Algorithm 1. This algorithm has time complexity $O(n)$. ■

For our evaluation, we use the most interesting case of the omniscient ISP and a fractional user.

3. ARCHITECTURE MODIFICATIONS

The proposed incentive scheme can be implemented with few a additions in the existing architecture of residential broadband networks. On the user side, it requires the addition of an interface for allowing user i to request an expansion ratio w_i in exchange for agreeing to limit the peak hour transmission rate to some level $U_i^{(n)} < U_i$ (via a rate controller at the OS level), U_i is the purchased nominal broadband access rate.

On the ISP side, a controller inside the DSLAM or at a higher layer should collect the w_i's before the beginning of the busy hour, compute the bids, and send acceptance signals carrying its offered reward rate $U_i^{(r)}$ for non-peak hours. For users that have gotten an offer, the DSLAM should monitor their transmission rate during the busy hour, and ensure compliance to the agreed rate $U_i^{(n)}$. At the end of the busy hour, the DSLAM should increase the maximum transmission rate to $U_i^{(r)}$ by tuning the transmission equipment (a standard capability in fiber to the home cards and existing ADSL equipment).

4. DATASETS

In this section, we present datasets used to evaluate our proposed scheme. We also present models for expansion ratios that users may demand from an ISP.

4.1 Access ISPs, a common transit provider

We have traffic load for all the PoPs of a large transit ISP (TR) that is the sole transit provider of several access ISPs, collectively serving more than 12 million ADSL users. TR connects to more than 200 other peer and client networks from all continents, for each of which, we have the nominal capacity, uplink and downlink volume of data transmitted over 5-minute intervals for several weeks of the first quarter of 2008.

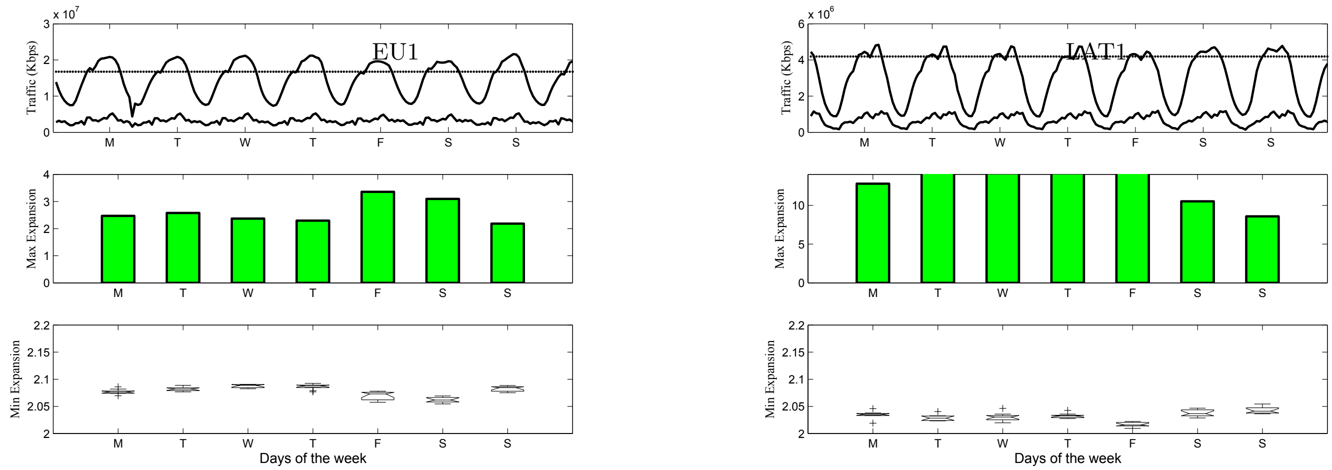

Figure 1: For each of the 2 customer provider links EU1, and LAT1: (top) aggregate and elastic traffic load under the pessimistic model for the classification of P2P traffic; (middle) maximum expansion under $\theta = 40\%$; (bottom) minimum expansion under $\theta = 40\%$, pessimistic classification, and omniscient fractional bidding with Pareto distributed user expansions w_i's of $\min(w_i) = 2$ and $E(w_i) = 5$.

In the above dataset, we pick two links, one connecting a large European access ISP to TR (EU1), and another connecting a large national Latin American ISP to TR (LAT1). The top row of each of the sub-plots of Fig. 1 plots the aggregate load on each one of these links for one week in March 2008. Interestingly, there are no significant weekend effects since these links cover mostly residential ADSL traffic and is weakly influenced by transition to and out of a weekend.

4.2 Per application classification of traffic

We do not have flow or application level information in our TR dataset above. Since these links serve a large number of end-users, we expect that the resulting classification will not be too different from other links with publicly available application level classification.

Using state of art tools on the WIDE backbone [5], a transpacific link that connects Japan to the US West Coast, we classify traffic by application including *"http"*[2]. After personal communication with maintainers of the WIDE repository, we consider two extreme cases: In the "pessimistic" case we assume that only *other* is P2P whereas in the "optimistic" case we assume that both *other* and *http* are P2P. We use the words pessimistic and optimistic in our ability to move elastic traffic to non-peak hours (with the optimistic case being easier). A recent study [11] of residential users from a large ISP in Europe also supports our claims and numbers.

An analysis of a 24-hour weekday trace for WIDE link from March 2008 for the pessimistic and optimistic cases revealed that the traffic volume for P2P followed a diurnal pattern for the optimistic and to a lesser extent for pessimistic case (figure not shown for lack of space). A very interesting observation was that percentage of P2P traffic did not vary substantially over the day (74-88% for the optimistic and 12-22% for the pessimistic case). We use these percentages in our links to identify elastic traffic.

4.3 Capacity of end-users

In order to create the inputs for our proposed bidding strategies, we need to know the amount of elastic traffic generated by each user. We assume that if a user is generating elastic traffic, e.g., P2P, then all the user capacity goes towards this. This is a reasonable assumption since the volume of P2P downloads far exceeds any other kind of traffic (i.e., web) for a single user. To assign the capacities for such elastic traffic generating users, we use the cumulative distribution function (cdf) of link capacities presented in [15].

4.4 Greediness of end-users

Measuring user greediness is nearly impossible without a large-scale user study. Such a goal, though in our future plan, is currently outside the scope of our study. To model human greediness ($w_i > 1$), we use the following synthetic mechanisms:

Uniformly distributed user greediness. In auction mechanisms it is often assumed that a player's valuation of an auctioned item follows a uniform distribution in the range $[w_{min}, w_{max}]$. In our case the "item" being auctioned is the shift of a user's elastic traffic into the non-peak hour and the single bidder is the ISP. We use $w_{min} = 2$ and will vary w_{max} (which we don't know) up until the point where the optimal bidding to keep the peak hour below θ (Sect. 2.2) will require providing to the users with accepted bids the maximum expansion ratio allowed by the threshold θ and the background load of the link (more on this in Sect. 5.1).

Pareto distributed user greediness. Inspired by the widely observed 80-20 rule [16] in the distribution of wealth[3] in free-market societies, we assume that a user's greediness follows the same profile as wealth. We draw expansion ratios from a Pareto distribution:

$$P\{w_i > w\} = k \cdot \frac{w_{min}^k}{w^{k+1}}$$

where w_{min} is the minimum greediness and k is the Pareto index which defines the moments of the distribution like its average value $E\{w_i\} = x_{min} \cdot k/(k-1)$.

5. EVALUATION

We present a trace-driven evaluation of the bidding algorithms of Sect 2.2 using the datasets from Sect. 4.

[2]Several P2P applications use http as transport protocol, or switch to it to bypass NATs and/or traffic engineering.

[3]It has been observed that in free-market societies 80% of the wealth is concentrated in 20% of the people.

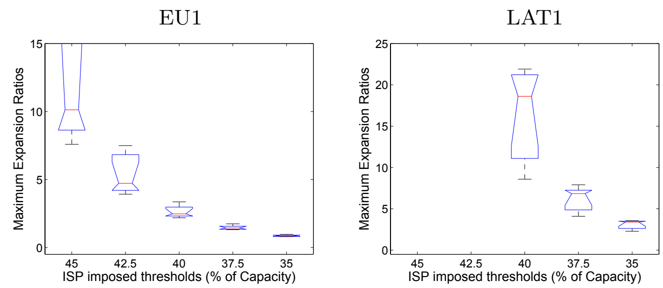

Figure 2: The figures show the maximum expansion ratios for θ values of 45, 42.5, 40, 37.5 and 35 % for both links, EU1 and LAT1.

5.1 Maximum expansion ratios

Given a time series x describing the traffic on one of the customer-provider links of Sect. 4 and a threshold θ on the utilization during the peak, for each day, we compute $P(\theta, x)$ the amount of traffic that must be removed from the peak to keep the utilization below the threshold. Similarly, we compute $V(\theta, x)$, the un-utilized capacity of the load valley that follows each daily peak. Thus, for $P(\theta, x) > 0$, the maximum aggregate expansion ratio that the ISP can provide is $w(\theta, x) = V(\theta, x)/P(\theta, x)$. In top row of Fig. 1, we plot the load time series x for both links, the elastic component, and a horizontal line indicating a threshold of 40%. The elastic component is obtained by multiplying $x(t)$ with $p_E(t)$, the percentage of traffic on the WIDE link that is classified as P2P at time t based on a pessimistic model. In the middle row of each sub-plot we plot the corresponding maximum aggregate expansion ratio $w(\theta, x)$ for day of the week. We observe that expansion ratios above 2 are available across all plots. Sub-plot LAT1 shows high expansion ratios on some days since only a very small percentage of traffic load is above the threshold of 40%. In sub-plot EU1, we see expected pattern of relatively low expansion ratios during weekdays with high traffic loads.

We emphasize that the maximum expansion available is independent of the models of user greediness that we use in the next sub-section.

Choice of θ, or an operational point plays a central role in the calculating the expansion ratios as we observe in Fig. 2. The time-series traffic at LAT1 is always below θ values of 42.5 and hence the Expansion ratio cannot be defined at those thresholds. From EU1 and LAT1, we observe that as θ goes down from 45% to 35%, the higher the traffic above the threshold the lower the expansion ratio and mean and variance.

5.2 Minimum expansion achieved by optimal bidding under peak utilization θ?

Having quantified the maximum expansion that an ISP can "spend" in incentivizing households, we show that through careful bidding, the ISP can move excess traffic with significantly lower expansion ratios. Thus, substantial amount of non-peak bandwidth may be used by the ISP for other purposes.

Input to the experiment. Given the traffic time series x, we obtain $\hat{x}$ the average aggregate rate during the peak hour of a day. We compute its elastic component $\hat{x} \cdot \hat{p}_E$ where $\hat{p}_E$ is the average value during the peak hour of $p_E(t)$, the percentage of elastic traffic (P2P) at time t, from Sect. 4.2 under the pessimistic model.

We model ADSL end-users by drawing values from the probability mass function (pmf) of active user capacity of Sect. 4.3. We stop when we have drawn $\hat{n}$ users whose aggregate capacity is exactly equal to the elastic traffic $\hat{x} \cdot \hat{p}_I$. We assign to each end-user a greediness w_i according to the Pareto and uniform models of Sect. 4.4. When drawing the capacity and the greediness from the aforementioned distributions, we do so independently for each random variable. We also experimented with positively and negatively correlated draws and obtained consistent results for each case. We do not present these results here for lack of space.

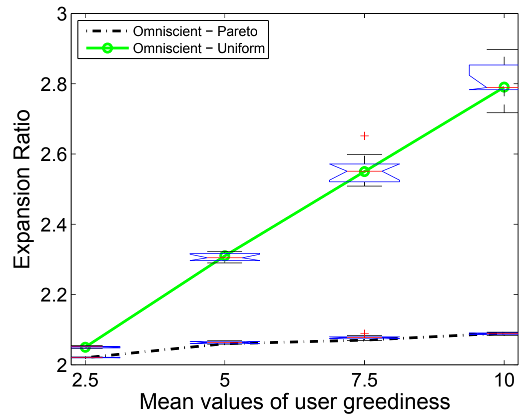

Figure 3: Effect of k on the expansion ratios. Observe that increasing maximum greediness increases the expansion ratios only marginally.

Minimum overhead for given θ and user model. We consider two distributions from which we draw user-greediness; uniform and Pareto distributions. In both cases, we use a minimum greediness of 2. Our first set of results are obtained for $\theta = 40\%$.

Uniform Distribution: For uniform distribution, we pick a minimum w_i and mean $E(w_i)$ and draw from the resulting distribution. Such a distribution may be characterized by saying that for any user, all values of a finite set of possible values are equally probable.

Pareto Distribution: For Pareto-distribution too, we pick a mean $E(w_i)$, and hence the resulting Pareto parameter k to draw from the distribution (e.g., for $E(w_i) = 5$, the corresponding Pareto parameter is $k = 5/3$). Having drawn the inputs for user greediness as described in the above two cases, we execute the optimal fractional omniscient bidding described in Algorithm 1 to obtain the corresponding minimum expansion ratios.

Results presented in the third row of each one of the sub-plots of Fig. 1 use a Pareto distribution with mean $E(w_i) = 5$. To remove the noise from the random drawing of end-user capacities and greediness in constructing the input, we repeat each experiment 10 times and report the median and the most important percentiles. We note that although LAT1 has high maximum expansion ratios during some days of the week, for the same day, the minimum expansion needed to satisfy the user demand is a little over 2 as an average.

Next, we compare results of expansion ratios by drawing user greediness from two distributions, uniform and Pareto, both with the same minimum and mean values. Figure. 3 depicts the median value and most important percentiles for the minimum expansion ratio under omniscient fractional bidding (all other biddings perform similarly) over a range of mean values. Although we allow the maximum values of the bids to increase, the expansion ratio for the Pareto model is almost flat. This occurs since the lowest w_i's are satisfied. Even if we sample from a uniform distribution, the increase in expansion ratio does not occur in the same ratio as the increase in maximum greediness.

In summary, assuming rational users, we presented a general mechanism for alleviating existing stress on network resources. By shifting delay-tolerant traffic in time, we reduce peak load while simultaneously increasing aggregate download volumes for users. Via careful bidding of expansion ratios, the ISP can incentivize users to move their elastic traffic to non-peak hours. We show that the ISP does not use its maximum available expansion, but a ratio substantially lower. It does this through a bidding strategy that satisfies the lowest w_i.

6. RELATED WORK

One of several measurement studies supporting the prevalence of P2P multimedia traffic on the Internet is [7]. There is extensive literature on usage-based pricing of network resources [3]. Despite the validity of the economic arguments presented in favor of usage-based pricing and against flat-rate, historic examples drawn from multiple areas of economic activity [13] seem to indicate that services like Internet access tend to converge towards very simple pricing schemes like the flat-rate. We use flat-rate pricing as the starting point of our work on which we offer incentives for users. Alternate pricing schemes have been proposed but these have been primarily ISP-based [14].

In a recent study [6] the authors infer median delay-tolerance of users (time from download to play) for podcasts to be of the order of tens of days. Thus, users may identify such content as being tolerant to delays for download. In our work, we assume delay-tolerance of users to be less than a day, a much stricter deadline. We also expect that incentives will work for other multimedia content like download of games, images and films. We see evidence of delay tolerance in bittorrent downloads [8]. We are not aware of work that has explored delay-tolerance of users by media type.

In our work, we assume that households honestly reveal their preference for expansion ratios. To the best of our knowledge, this is the first work to explicitly deal with an incentive scheme built into flat-rate pricing.

7. CONCLUSION

In this work, assuming rational usage in a home, we presented a general mechanism for alleviating existing stress on network resources. By shifting delay-tolerant traffic in time, we reduce peak load while simultaneously increasing aggregate download volumes for users. We also showed how such schemes can be made easy-to-adopt by keeping them flat-rate compatible and provided architectural details for enabling them. We modeled ISPs and users and characterized the complexity of computing the optimal incentive. We provided an extensive empirical validation of our scheme by doing a detailed trace-driven analysis of the gains from time-shifting traffic using two different and widely accepted models of users' utilities. In future work we propose to obtain more granular data and analyze it using additional models[4] of users' utilities. But the ultimate test of the worth of our ideas lies in their adoption by end users and ISPs. To this end we are building such a system for deployment.

8. ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their useful comments in improving the paper.

9. REFERENCES

[1] Xbox live and netflix. At http://www.xbox.com/en-US/live/netflix/default.htm.

[2] Demetris Antoniades, Evangelos P. Markatos, and Constantine Dovrolis. One-click hosting services: a file-sharing hideout. In *Proc. of IMC'09*.

[3] C. Courcoubetis and R. Weber. *Pricing Communication Networks: Economics, Technology and Modelling*. Wiley, 2003.

[4] J. de Vries. *The Industrious Revolution: Consumer Behavior and the Household Economy, 1650 to the Present*. Cambridge University Press, 2008.

[5] MAWI Working Group. Packet traces from WIDE backbone. http://tracer.csl.sony.co.jp/mawi/.

[6] D. Gunawardena, T. Karagiannis, A. Proutiére, and M. Vojnovic. Characterizing podcast services: publishing, usage, and dissemination. In *In Proc. of IMC*, pages 209–222, Chicago, IL, 2009.

[7] L. Guo, S Chen, Z Xiao, E Tan, X Ding, and X Zhang. Measurements, analysis, and modeling of bittorrent-like systems. In *Proc. of IMC*, 2005.

[8] M. Izal, G. Urvoy-Keller, E. Biersack, P. Felber, A. Al Hamra, and L. Garces-Erice. Dissecting bittorrent: Five months in a torrent's lifetime. In *Proc. of PAM*, 2004.

[9] T. Karagiannis, E. Athanasopoulos, K. Gkantsidis, and P. Key. Homemaestro: Order from chaos in home networks. Technical report msr-tr2008-84, Microsoft Research, 2008.

[10] N. Laoutaris and P. Rodriguez. Good things come to those who (can) wait or how to handle delay tolerant traffic and make peace on the Internet. In *Proc. of ACM HotNets-VII*, 2008.

[11] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On dominant characteristics of residential broadband internet traffic. In *Proceedings of the IMC*, pages 90–102, 2009.

[12] MSNBC. Comcast blocks some internet traffic, 2007. http://www.msnbc.msn.com/id/21376597/.

[13] A. Odlyzko. Internet pricing and the history of communications. *Computer Networks*, 2001.

[14] Andrew M. Odlyzko. Paris metro pricing for the internet. In *ACM Conference on Electronic Commerce*, pages 140–147, 1999.

[15] M. Piatek, T. Isdal, T. Anderson, A. Krishnamurthy, and A. Venkataramani. Do incentives build robustness in BitTorrent? In *Proc. of NSDI*, Cambridge, MA, 2007.

[16] William J. Reed. The Pareto, Zipf and other power laws. *Economics Letters*, 74(1):15–19, 2001.

[4]We note the limitations of the use of Pareto distribution. However, the penalty of using the uniform distribution did not increase in the same factor as the increase in maximum greediness.