Evaluating Tradeoffs of Congestion Pricing for Voice Calls

Jimmy S. Shih
University of California at Berkeley
Room 473 Soda Hall #1776
Berkeley, CA 94720-1776
jshih@cs.berkeley.edu

Randy H. Katz
University of California at Berkeley
Room 637 Soda Hall #1776
Berkeley, CA 94720-1776
randy@cs.berkeley.edu

Keywords: Resource allocation, dynamic pricing, user study, voice over IP, computer telephony service.

Abstract

Congestion pricing, varying prices according to the load, can more efficiently allocate scarce resources by modifying user behaviors. However, it presents operators with a tradeoff between system performance and user satisfaction. To evaluate this tradeoff, we propose a methodology that combines small-scale user experiments with simulations, and apply it to voice calls. We offered an IP-telephony service to 100 users over one year where each user is allocated a fixed number of tokens a week. Based on extensive experimentation, we found that the users are responsive to occasional price increases and that they can be enticed to accept congestion pricing. Using the results from the experiments, we developed a user behavioral model to simulate the effects of congestion pricing for larger user populations under the same token constraint. We proposed a set of thumb-rules operators can use to set the parameters for managing congestion pricing. By setting these to appropriate values, we estimated that congestion pricing can reduce call blocking rate by 60% or save provisioning by 20% while causing users to experience price changes in 4% of their usages. Our estimates strongly depend on our derived user model, thus we further verified the model by having real users using the service along with simulated users. Our findings indicate that congestion pricing can be very effective for voice calls because it only needs to be applied occasionally during high loads, and that users are tolerant and responsive to these occasional price increases.

1. INTRODUCTION

Congestion pricing, varying prices according to the load, can more efficiently allocate scarce resources by modifying user behaviors. It can entice users to conserve resources or defer usages during congested periods. For voice calls, it can encourage users to talk shorter or talk at another time. By involving users in resource allocation decisions, it offers operators an extra dimension of flexibility in making tradeoffs between system performance and user satisfaction. It can improve system performance by reducing congestion, supporting more users with a given capacity, and saving money in terms of the needed capacity. Operators can then pass along some of these benefits to users by reducing prices or congestion level. On the other hand, it can reduce user satisfaction by interrupting users with announcements of price changes and causing users not to be able to predict their costs ahead of the time.

The state-of-the-art on congestion pricing consists of papers [1-3] that analyze its advantages and disadvantages. Using simulations, other papers [4-8] have shown that congestion pricing can potentially be better or worse than flat-rate pricing. These simulation results are limited to qualifying its benefits and drawbacks because they are strongly dependent on how they model user behavioral response to price changes. Without a realistic user model based on user experimentation, it is difficult to quantify the tradeoffs of congestion pricing.

Our goal is to estimate how well congestion pricing can work in a system designed to support a very large community of users. However, it is infeasible to perform a study involving many users. Thus we propose a methodology that combines small-scale user experiments with simulations to understand the tradeoffs of congestion pricing. We use the user experiments to understand how users would behave under congestion pricing. In particular, whether they would react to dynamic prices, how sensitive they are to different price increases, and whether they would accept congestion pricing. Then using a user model derived from the user experiments, we use the simulations to understand the tradeoffs an operator would face when there are many users, and how it should set the appropriate parameters for managing congestion pricing. Afterwards, we combine the user experiments with the simulations to further verify the user model. We do so by having users reacting to price changes set by an operator who is responding to the load and the reactions of many simulated users. This methodology allows us to use the observations from a small user community to model the behaviors of a larger community.

We apply the above methodology to an IP-telephony service where users can use their computers or phones to make and receive telephone calls. The service is ideal for applying congestion pricing because the Internet-to-PSTN

gateway it uses is a natural bottleneck due to a limited number of phone lines to the PSTN. It is also ideal because we can easily perform load monitoring, admission control, accounting, and price setting at the gateway.

Using the methodology, we found that congestion pricing can be very effective for voice calls because prices only need to be changed occasionally and that users would willingly respond to these occasional price increases. In reaching this conclusion, we present an effective way of applying congestion pricing to users, a user model for modeling user reaction to price changes, a set of rules for operators to manage congestion pricing, and an estimate of the benefits and drawbacks of congestion pricing.

For the rest of this paper, we first discuss the related work in section 2, and our methodology in section 3. Afterwards we describe the user experiments in section 4, and our simulation results in section 5. Finally we conclude with our findings about congestion pricing in section 6.

2. RELATED WORK

MacKie-Mason and Varian [1] and MacKie-Mason, Murphy, and Murphy [2] advocate the use of congestion pricing for allocating scarce resources. They point out that congestion pricing can reduce congestion, better utilize available resources, send correct signals about expansion, and place no artificial limit on usages. However they also mention that congestion pricing requires a more complicated accounting mechanism, offers less predictable prices, and needs users to be in the control loop. MacKie-Mason and Varian summarize their case for congestion pricing by explaining that fluctuation in either prices or delay is unavoidable in a congested network, thus congestion pricing at least allows operators to make a tradeoff between the two. Fishburn and Odlyzko [3], on the other hand, argue that flat-rate pricing is better than congestion pricing when technology improves rapidly because it only requires a little bit more resources to gradually improve quality for everyone.

Though there are many research efforts which argue for and against the use of congestion pricing, there are very few works which actually perform pricing experiments with real users. Klausz et al. [9] conducted simulated games where users use limited tokens to participate in auctions of modem usages. They used Nth-price auction to encourage users to use the modems at off-peak hours. They found that prices set in these simulated auctions can increase utilization, reduce blocking rate, and increase users' utility functions. The INDEX project [10] performed a detailed user study to measure user demand for Internet bandwidth. They offered users several bandwidth choices with fixed prices and allowed them to change their choices at any time. They found that different pricing policies, like by volume or by connect time, can significantly affect user demand. For each pricing policy, they found that the demand is very sensitive to prices and that the difference among users is persistent and large. Our work is similar to Klause et al. [9] in that we also use a token system as our budget constraint, however, we use a real service and dynamically adjust prices during usages instead of using simulated auctions and letting users indicate the most they are willing to pay. We differ from the INDEX project in that we measure the effectiveness of changing prices in the middle of usage whereas the INDEX project measured the user demand for different bandwidth when the prices are fixed.

There are many works [4-8] which use simulations to compare congestion pricing with flat-rate pricing. Cocchi et al. [4] found that it is possible to set prices so that everyone is better off because the performance penalty received for requesting a less-than-optimal service is offset by the reduced price of the service. Peha [5] found that in terms of the values of the transmitted data, congestion pricing can be better than allocating everyone equal share of the bandwidth. Murphy et al. [6] used simulations to show that congestion pricing can allow a network to carry more traffic and that users would value the traffic more. Finally, Wang and Schulzrinne [7] showed that congestion pricing can achieve stable and fair distribution of bandwidth between users according to their demand elasticity. However, Paschalidis and Tsitsiklis [8] showed through their simulations that a static pricing policy like time-of-day pricing is almost as effective as congestion pricing. These simulation results [4-8] strongly depend on how they model user reaction to price changes. They are limited to qualifying the benefits and drawbacks of congestion pricing because without a realistic user model, it is difficult to quantify the tradeoffs. We use the data from our user experiments to formulate a user model so that we can perform large-scale simulations to study congestion pricing.

3. METHODOLOGY

It is difficult to perform a large-scale user study involving many users, thus we propose a four-steps methodology, see Figure 1, for evaluating congestion pricing. The first step develops a testbed to demonstrate a system using congestion pricing. The second step performs user experiments with a small number of users to observe user reaction to price changes. The third step uses a user model based on the user experiments to perform large-scale simulations to understand the management and the tradeoffs of congestion pricing. Finally, the fourth step uses the simulation results to verify the user reaction to price changes by having real users using the system along with a large number of simulated users.

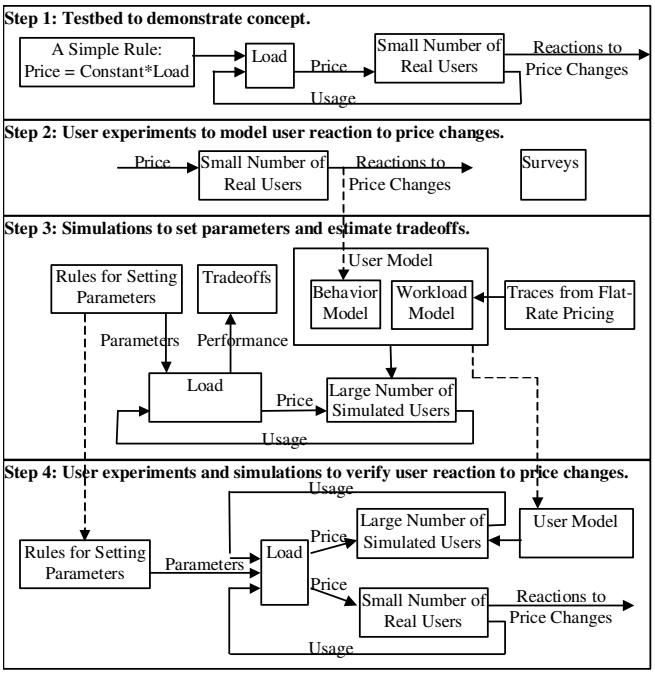

Figure 1. A four-steps methodology for evaluating congestion pricing.

We applied the methodology to voice calls. In the first step, we prototyped an IP-telephony service using a H.323 gateway [11] as the Internet-to-PSTN gateway. The gateway has a Primary Rate ISDN connection to the PSTN capable of supporting 23 simultaneous calls. We then developed a web interface and an Interactive Voice Response system for the gateway so that users can use their computers or phones to make and receive phone calls. When the users are on their computers, they can see the current prices on the web interface and when they are on their phones, they will hear short announcements of price changes. See reference [12] for more details about the service.

In the second step, we deployed the service to users to conduct pricing experiments. We did not charge users real money[1], instead we limited each to a certain number of tokens a week. In the service, we occasionally adjusted the token rates during their phone calls to observe how they would react to such changes. We experimented with different ways of applying congestion pricing, measured how sensitive users are to price increases, and surveyed users about their acceptances to congestion pricing.

After understanding user reaction to price changes, in the third step, we performed simulations to measure the effectiveness of congestion pricing when there are many users using the service under the same token constraint. In the simulations, we used traces from flat-rate pricing to generate a user's workload model and observations from the user experiments to model his/her behavior to price changes. We first determined the optimal parameter settings an operator would use when applying congestion pricing to thousands of users. Then using the appropriate settings, we estimated the potential benefits and drawbacks of congestion pricing.

Our estimates strongly depend on our user model, so in the fourth step, we verified it by having a new group of users using the service along with simulated users. The simulated users would make calls and react to price changes according to the model, the operator of the service would use the appropriate parameters to manage congestion pricing, and we would observe real users' reactions to price changes under this setting.

4. USER EXPERIMENTS

4.1 Goals

The goals of our user experiments are to understand how to apply congestion pricing to users and how users would react to price changes. For applying congestion pricing, we need to know how often prices can change and whether changing prices in the first few minutes of a call would be effective. For reacting to price changes, we want to measure the effects of different price increases on causing users to terminate their calls earlier and survey users about their acceptances to congestion pricing. By understanding how to apply congestion pricing and how users would react to it, we can later better model user behavior.

4.2 Experimental Setup

We offered the IP-telephony service to 100 dormitory students, of various majors, for over one year. We limited each student to 1000 tokens a week, with unused tokens disappear at the end of the week, and charged a certain token rate a minute. We chose 1000 tokens and the charging rate so that an average user will run out of tokens by the end of the week. We randomly divided our users into two groups and experimented with a different pricing policy each week, see Table 1. We first experimented with congestion pricing (3/5/01 and 3/12/01) where price varies according to the load. We set the price to be the number of simultaneous calls multiplied by a constant. In these two weeks, we had up to 5 simultaneous calls. We then experimented with how often price can change and how much the price should change each time (4/16/01, 4/23/01, and 4/30/01). In these experiments, we did not want to change the price frequently and would like to charge our users the same initial rate most of the time. Occasionally, we would increase the price from the initial rate and then decrease it back again to measure the effectiveness of different price increases from the initial rate. In between some of the experiments, we used a flat-rate pricing for comparison. Each week, we had about 45 active users making 250 calls totaling 2500 minutes.

[1] Charging real money would have complicated the approval process for conducting our user experiments.

Table 1. Pricing policies.

Base case: Initial price of 10 tokens/min. Price increases by 10 tokens and then decreases by 10 tokens. Price can change from one minute to the next, but changes on the average once every 10 minutes.		
Week	Group 1 policy	Group 2 policy
2/26/01	Flat-rate of 10 tokens/min.	Flat-rate of 10 tokens/min.
3/5/01	Price varies according to the load, equals **10 times** the number of simultaneous calls.	Price varies according to the load, equals **10 times** the number of simultaneous calls.
3/12/01	Price varies according to the load, equals **5 times** the number of simultaneous calls.	Price varies according to the load, equals **5 times** the number of simultaneous calls.
4/16/01	Base case.	Base case except price can only increase, by 5 tokens each time, with a maximum price of 25 tokens/min.
4/23/01	Base case except price can change at most once every **3 minutes**.	Base case except price can change at most once every **5 minutes**.
4/30/01	Base case except price increases by **5 tokens** and then decreases by 5 tokens. Price can change at most once every **3 minutes**.	Base case except price increases by **15 tokens** and then decreases by 15 tokens. Price can change at most once every **3 minutes**.

4.3 Results

The main result from the experiments is that rapid price changes are ineffective for changing user behavior. When we allowed the price to change from one minute to the next according to the load, we found that the price changes did not cause users to terminate their calls much earlier. Based on 1140 calls under flat-rate pricing, we expect that 5.8%, with a standard error of 0.4%, of the active calls to terminate each minute. In Figure 2, we observe that the percentage of the active calls terminating after a price change remains around 5.8%. Based on a follow-up survey, users indicate that even though they pay attention to the price changes, they feel that after a price increase, they can just wait till the next minute and hope that the price would decrease. We then experimented with making the price change at most once every three minutes. We artificially increased the price by a certain increment from an initial rate and then decreased it back. We found that depending on the price increment, we can get different percentage of the active calls to terminate right away, see Figure 3. Thus based on the user experiments and the follow-up survey, it is important that when the price changes, it does not change from one minute to the next so that users can better predict the cost of not terminating their calls earlier.

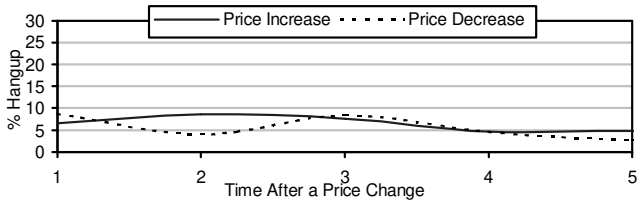

Figure 2. Price can change from one minute to the next[2].

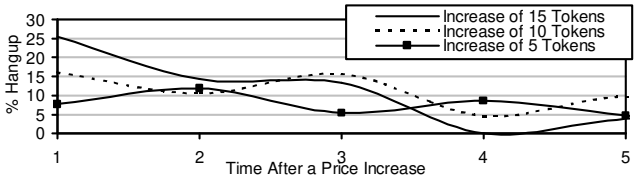

Figure 3. Price can change at most once every three minutes[3]

Our next result is that the first few minutes of a call are very important to users. Under flat-rate pricing, as call duration increases, the percentage of calls terminating each minute quickly decreases to a constant rate, see Figure 4. When we increased the price on the 3rd minute of a call, we found that we could not get users to terminate their calls more when compared to flat-rate pricing. Calls that terminate quickly after a price increase in the beginning of the 3rd minute would register as calls lasting 4 minutes. However, when we increased the price on the 5th minute of a call, we found that we could start to get more people to terminate their calls early, see Figure 4 for calls that last only 6 minutes. Thus we believe that inorder for congestion pricing to be effective, it should not be applied in the first three minutes of a call.

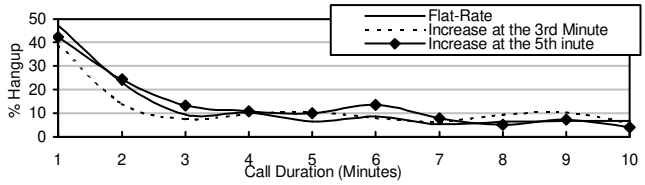

Figure 4. Price increase at different minute of a call[4].

There are three other interesting results we found from our experiments when the price changes at most once every three minutes. For the price decreases associate with Figure 3, we found that we still get close to 5.8% of the calls to terminate the next minute instead of a percentage like 1% or 2%, see Figure 5. Thus users really do not try to talk longer after a price decrease. Based on the follow-up survey, users indicate that they feel there is no need to talk longer after a

[2] Data fluctuates because it contains only 137 samples of price increase and 135 samples of price decrease. These price changes were not frequent, the price increases occurred in 2.4% of the minutes used by the users and the price decreases occurred in another 2.4%.

[3] Data fluctuates because it contains only 47 samples of price increase of 15 tokens, 68 samples of price increase of 10 tokens, and 91 samples of price increase of 5 tokens. These price increases occurred in 6.2% of the minutes used by the users and their associated price decreases occurred in another 2.6%.

[4] Data contains1140 calls for flat-rate pricing, 128 calls for increase at the 3rd minute, and 243 calls for increase at the 5th minute. Calls terminating in 1 minute include busy calls and no answered calls.

price decrease. Second, we found that the higher the price increase is, the more users would hangup and call back within a minute to obtain the lower initial rate, see Figure 6. Thus this behavior slightly reduces the effectiveness of congestion pricing. Third, we found that congestion pricing cause the peak period to shorten from that of flat-rate pricing, see Figure 7[5]. Thus congestion pricing can also reduce congestion by enticing users to defer some of their peak usages to another time.

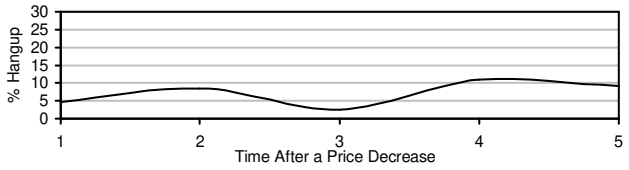

Figure 5. Percentage hangup after a price decrease[6].

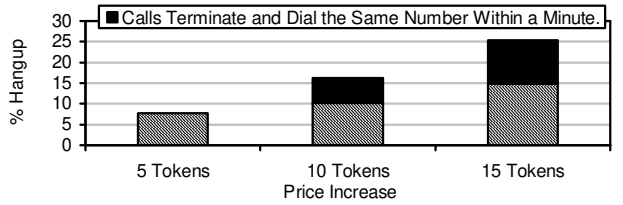

Figure 6. Breakdown of calls that terminate within a minute of a price increase.

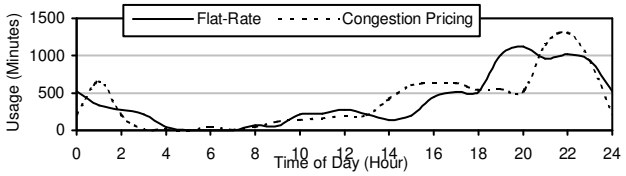

Figure 7. Flat-rate and congestion pricing calling pattern[7].

4.4 Survey

After the experiments, we used another follow-up survey to confirm the results of our observations[8]. In the survey, we want to find out more about users' acceptances to dynamic prices, financial incentives for choosing congestion pricing, and stated behaviors to price changes. For the user acceptances, we first asked users whether they like congestion pricing. Most of the users did not like it because of the price-change announcements interrupting their calls. However, when we pointed out that congestion pricing can reduce call blocking rate or make prices cheaper, more users liked it. We also asked the users how they would like congestion pricing if we use beeps, instead of recorded messages, to indicate the current prices. For example, 1 beep for 10 tokens, 2 beeps for 20 tokens, etc. We found that more users would accept congestion pricing if the user interface is not as disruptive as inserting recorded messages in the middle of phone calls. See Table 2 for the questions and their average scores.

Table 2. Survey regard user acceptances.

From a score of 1-5 (5:like it a lot, 4:like it, 3:OK, 2:dislike it, 1:dislike it a lot),	Average score (standard error)
how do you like congestion pricing?	2.61 (.19)
what if it can reduce the chance that your calls might be blocked?	3.04 (.20)
what if it can make it cheaper for you to use the service?	3.91 (.16)
what if there is a less disruptive way to announcing price changes like using beeps?	4.70 (.12)

For the financial incentives for choosing congestion pricing, we want to know how much discount in tokens we need to offer to users for them to choose congestion pricing over flat-rate pricing. In the survey, we asked the users to choose between a congestion pricing policy of certain average rate or a flat-rate pricing of certain rate. See Table 3 for the list of the choices. As we made the rate under flat-rate pricing higher, we could easily entice more users to choose congestion pricing over flat-rate pricing. This indicates that the users would be willing to use congestion pricing if given a small discount.

Table 3. Survey about financial incentives.

Do you prefer congestion pricing with an average rate of 12.5 tokens/min (80% 10 tokens, 15% 20 tokens, and 5% 30 tokens) or	0 for congestion pricing and 1 for flat-rate pricing. Average score (standard error)
flat-rate of 12.5 tokens/min?	0.89 (.07)
flat-rate of 15 tokens/min?	0.37 (.11)
flat-rate of 20 tokens/min?	0.11 (.07)
flat-rate of 25 tokens/min?	0.05 (0.5)

Finally, for the stated behaviors to price changes, we questioned our users how they would react to a price change. Most users answered that they would talk shorter when the price increases, but not longer when the price decreases. However, our users stated that they would behave about the same whether the price changes from one minute to the next or changes at most once every three minutes. Thus from using surveys, it is difficult to distinguish the effects of the frequency of price changes. See Table 4 for the questions and their average scores.

Table 4. Survey of stated behaviors.

When the price can change	1 for yes and 0 for no. Average score (standard error)
from one minute to the next, would a price increase affect your behavior?	0.36 (.13)
from one minute to the next, would a price decrease affect your behavior?	0.21 (.11)
at most once every 3 minutes, would a price increase affect your behavior?	0.32 (.12)
at most once every 3 minutes, would a price decrease affect your behavior?	0.14 (.10)

5. SIMULATIONS

5.1 Goals & Metrics

There are two questions we want to know through simulations. First, how should a telecommunication operator apply congestion pricing when there are many users. In

[5] In Figure 7, the peak under congestion pricing is higher than the peak under flat-rate pricing assuming both pricing policies handle the same number of minutes. However, based on our later simulation results, we estimate that congestion pricing can reduce the peak load by 20%, thus making both peaks the same.

[6] Data contains 86 price decrease.

[7] Data contains 9,299 minutes of calls under flat-rate pricing and 3,347 minutes of calls under congestion pricing. The data from the congestion pricing is scaled to the data of the flat-rate pricing for comparison.

[8] We received 23 responses.

particular, how should the operator set the following four parameters.

- *Capacity* – how many phone lines at the Internet-to-PSTN gateway to purchase.
- *Threshold* – at what load level should he/she start applying congestion pricing.
- *Interval* – how quickly should prices be allowed to change.
- *Init* – how long to wait on new calls before applying congestion pricing to them.

Second, using the appropriate parameter values, what are the potential benefits and drawbacks of congestion pricing. For the benefits, the operator would like to know the expected improvements in call blocking rate or provisioning. For the drawbacks, the operators would like to know how often users would encounter a price change. There is always a tension between system performance and user satisfaction when applying congestion pricing. Flat-rate pricing can be viewed as one extreme where users are very satisfied because prices do not vary but that the operator needs to provision more capacity or allow the system to be congested. Thus the operator needs to make a tradeoff between system performance and user satisfaction when applying congestion pricing.

5.2 Modeling

There are two parts to modeling a user's behavior. First, when would a user make a call and second, how would he/she react to a price change. For both parts, we model the user when she/he is limited to 1000 tokens a week so that we can take advantage of our experimental results. For a user's calling patterns, we have traces of 1140 calls totaling 9299 minutes made under flat-rate pricing with the token constraint. These calls would be made by 175 users in a week. Thus for modeling a user's calling pattern, we have traces of the number of calls he/she would make in a week, see Figure 8, when these calls would be made, see Figure 9, and how long these calls would last, see Figure 10. Thus we can use these traces to simulate the workload of a large number of users. For example, Figure 11 shows a load made in a week by 10,000 simulated users making 64,510 calls totaling 516,235 minutes. The load in the Figure exhibits periods of high utilization followed by low utilization.

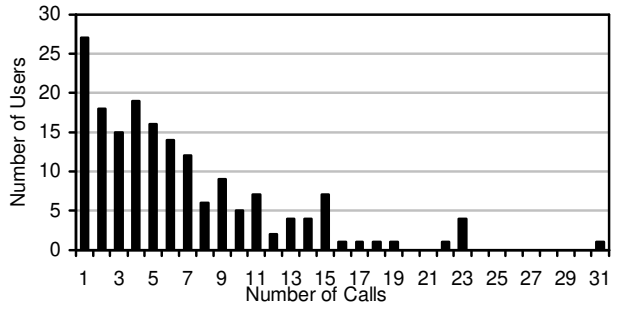

Figure 8. Traces of the number of calls a user makes a week.

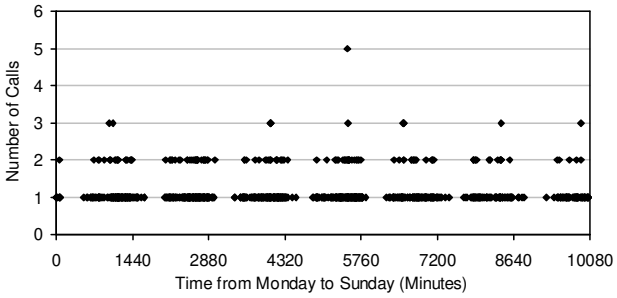

Figure 9. Traces of call starting times.

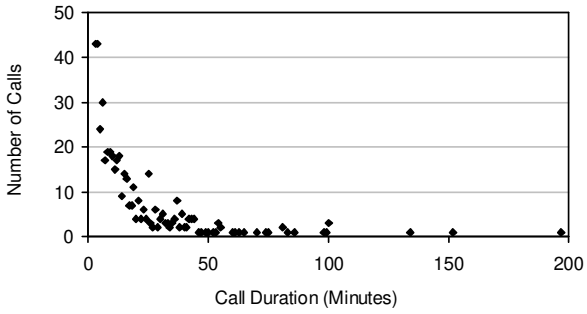

Figure 10. Traces of call durations[9].

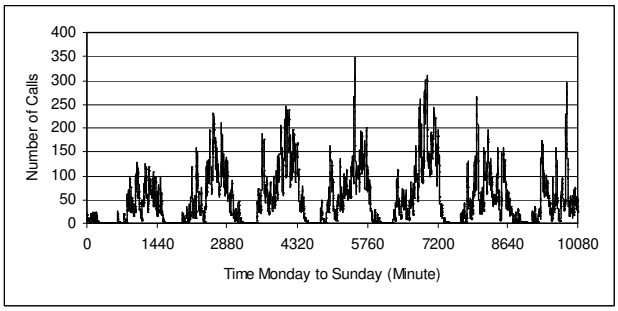

Figure 11. One week of calling pattern by 10,000 simulated users.

The second part of the user model is how a user would react to a price change. Using a simple model, we assume that the user would terminate his/her call with a certain probability after a price increase if prices do not change rapidly nor in the first few minutes of a call. However, the user would not talk longer after a price decrease. We also assume that there is a minimum and a maximum rate for the price and that the price increases or decreases by the same increment each time. We assume that the user behavior is more dependent on the price increment than on the price level because we are modeling an environment where the price would be a certain rate most of the time and would only change occasionally. We use this model because we were able to verify that real users would indeed terminate their calls early under such situation.

We model the operator by assuming that he/she has a limited *capacity* of phone lines to the PSTN. When all the phone lines are used, the operator would need to block new calls. However, the operator can increase prices when the

[9] Not shown are calls lasting one minute (539 calls) and two minutes (139 calls).

load is above a *threshold* value and decrease the prices when the load is below it. Through simulations, we measure the overall call blocking rate and the percentage of the minutes used by users that would encounter a price change.

5.3 Simulation Setup

We simulated the effects of congestion pricing when there is a fixed load of 5,000, 10,000, 50,000, and 100,000 users. In the simulations, we varied the four parameters, *capacity*, *threshold*, *interval*, and *init*. For all the following simulation results, we set the minimum rate to 10 tokens, the maximum rate to 40 tokens, the price increment to 10 tokens, and the probability of a call terminating after a price increase to 15%[10].

5.4 Results

The most important parameter we found for affecting performance is the *threshold* value. We found that an operator should apply congestion pricing when the load is about 90% of the *capacity*. Figure 12 shows the simulation results when there are 10,000 users, the *interval* is set to three minutes, and the *init* is also set to three minutes. If we use a lower *threshold* value, we would not gain much in terms of reduction in call blocking rate, however, if we use a higher *threshold* value, then the call blocking rate would quickly increase. Thus by setting the *threshold* value to 90% of the *capacity*, the operator would not need to change prices frequently and would still obtain the lowest call blocking rate. We also found that there is a great tension between call blocking rate and price announcement rate. As one aggressively applies congestion pricing, the call blocking rate would reduce but the price announcement rate would increase. Figure 13 graphs the price announcement rate as the *threshold* value changes under the same settings. At the time when the call blocking rate increases sharply, the price announcement rate would decrease sharply.

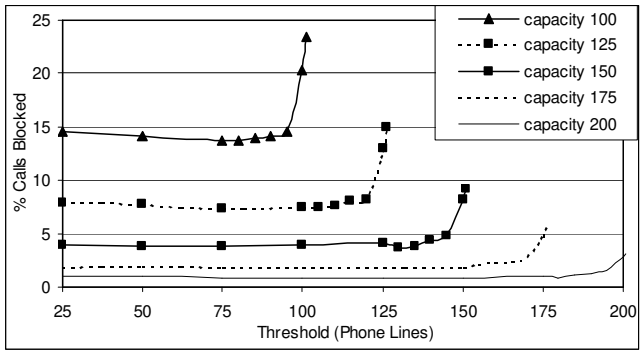

Figure 12. Call blocking rate for different *threshold* values.

[10] We also performed simulations for 12% and 20%.

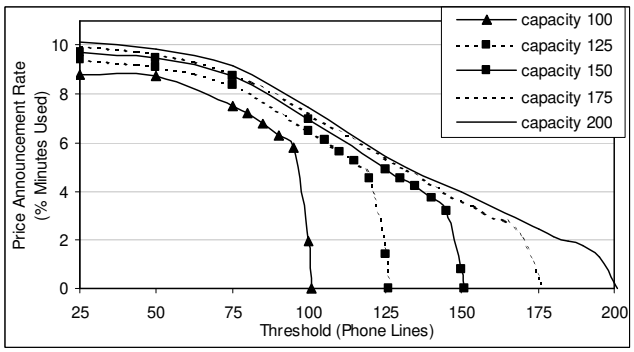

Figure 13. Price announcement rate for different *threshold* values.

We then vary the *interval* and the *init* parameters to observe their effects on performance. In Figure 14 and Figure 15, we show the call blocking rate and the price announcement rate when there are 10,000 users and the *capacity* is set to 150 phone lines. We found that when the *threshold* is set to 90% of the *capacity*, as we change the *interval* from 4 minutes to 2 minutes and also change the *init* from 4 minutes to 2 minutes, we can gradually decrease the call blocking rate from 4.8% to 3.4%, but at the same time gradually increase the price announcement rate from 3.6% to 5.0%. From our simulations, *interval* is more effective than *init* because it is applied all the time instead of just once in the beginning of a call. Changing these two parameters can be useful when it is difficult to quickly adjust the *capacity* and the load suddenly changes.

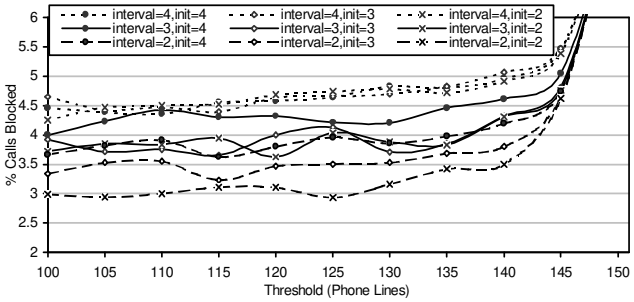

Figure 14. Call blocking rate as *interval* and *init* change.

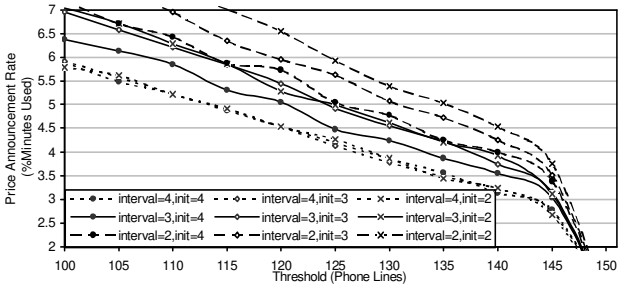

Figure 15. Price announcement rate as *interval* and *init* change.

Finally, the operator should choose a *capacity* to meet its call blocking rate and price announcement rate requirements. In Figure 16 and Figure 17, we graph the call blocking rate and the price announcement rate that the operator would expect when there are 10,000 users, the *threshold* is set to 90% of the *capacity*, the *interval* is set to

three minutes, and the *init* is also set to three minutes. For example, if an operator is willing to tolerate 3.8% call blocking rate and 4.2% price announcement rate[11], then he/she can purchase 150 phone lines to support the load. The operator can change the *init* and the *interval* values to obtain a slightly different tradeoff.

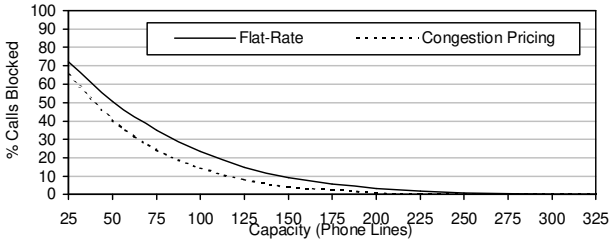

Figure 16. Call blocking rate under flat-rate and congestion pricing.

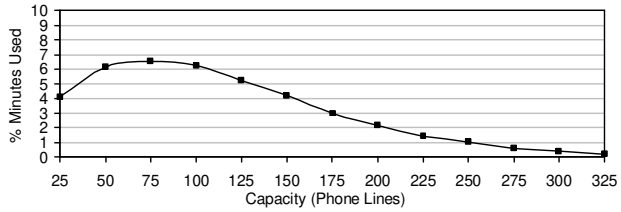

Figure 17. Price announcement rate under congestion pricing.

Figure 16 and Figure 17 also show the expected benefits and drawbacks of congestion pricing over flat-rate pricing. For example, when there are 10,000 users and the operator has 150 lines, he/she can expect the call blocking rate to reduce by 60% (from 9.2% to 3.8%) or expect to save provisioning by 20% (from 150 lines to 120 lines). In terms of the price announcement rate, he/she can expect that the users would experience price changes in 4.2% of their usages. The benefits and drawbacks are limited because there are only a few congested periods in a given day, price announcements can only be applied to calls longer than three minutes and have not had their prices changed in the last three minutes. Thus by using the appropriate parameters, congestion pricing will not cause many price changes, and at the same time would only be able to improve performance by a certain extent.

5.5 Verification

Our simulation results strongly depend on our user model that was derived when we randomly increase the price from an initial rate and then decrease it back. So we would like to further verify it by conducting user experiments where users would react to price changes under a more realistic setting. In particular, we would like the price changes to be caused by the actions of a large number of simulated users. These simulated users would use the user model to make calls and react to price changes. As the operator, we would apply the appropriate congestion pricing parameters when managing many users and allow prices to slowly vary between a minimum rate and a maximum rate. Then we would observe if the real users still behave as expected in the model.

In our verification, we had 100 new dormitory students using the service along with 10,000 simulated users. The minimum rate is 10 tokens, the maximum rate is 40 tokens, and the price increment is 10 tokens. We assumed that 15% of the time, a simulated user would terminate his/her call after a price increase. We simulated that the operator has a *capacity* of 150 phone lines, sets the *threshold* at 135 phone lines (90% of the *capacity*), sets the *interval* to three minutes, and sets the *init* value to three minutes. We randomly divided the dormitory students into two groups, a test group and a control group, for performing the experiments, see Table 5 below. Each week, we had about 85 active users making 600 calls totaling 6000 minutes. In these experiments, we observed the percentage of calls terminating after different price increases.

Table 5. Pricing policies for verifying the user model.

Base case: Initial rate of 10 tokens/min. Minimum rate of 10 tokens/min. Maximum rate of 40 tokens/min. Price changes by **10 tokens** each time. Price changes at most once every 3 minutes.		
Week	Group 1 policy	Group 2 policy
10/1/01	Base case.	Base case.
10/29/01	Base case except price changes by **5 tokens** each time.	Base case.
11/5/01	Base case except price change by **15 token**s each time.	Base case.

Our results further confirmed our user model that we can easily entice users to terminate their calls after a price increase if prices change slowly, at most once every 3 minutes, and infrequently, at most 4% of the time. As shown in Figure 19, Group 1 would terminate their calls more after a price increase, more so with a higher price increment. During the three-weeks experiment, the control group, Group 2, consistently terminates its calls around 10% after a price increase of 10 tokens. Both groups did not talk longer after a price decrease. During the experiments, users only encountered price announcements, price increases and price decreases, in about 4% of their usages. In Figure 20, we also graph the calling pattern during the experiments and compare it to flat-rate pricing. The Figure further confirms that congestion pricing can also encourage users to defer some of their usages from peak times to off-peak times[12].

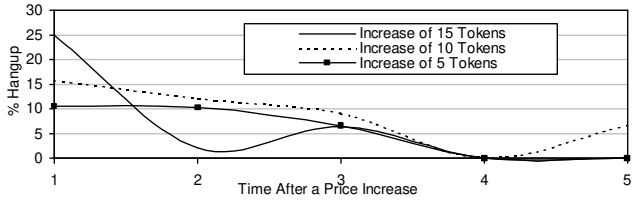

Figure 18. Group 1 reaction to different price increases.

[11] 3.0% are price increases and 1.2% are price decreases.

[12] In Figure 19 assumes both congestion pricing and flat-rate pricing would carry the same number of minutes. However, as shown in our simulations, we estimate that congestion pricing would be able to carry 20% less traffic during peak times than flat-rate pricing.

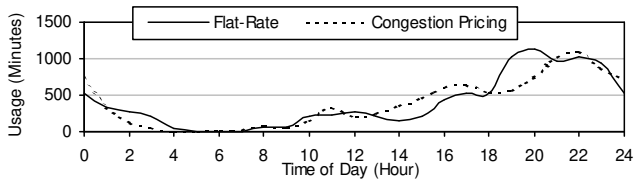

Figure 19. Flat-rate and congestion pricing calling pattern[13].

6. CONCLUSION

We would like to evaluate the tradeoffs of congestion pricing between system performance and user satisfaction in systems designed to serve many users. However, it is difficult to conduct a large-scale user study. Thus we propose a methodology that combines small-scale user experiments with simulations for evaluating the tradeoffs under high loads. We applied the methodology to voice calls by offering an IP-telephony service to 100 students in the dormitories for one year. We gave each student a certain number of tokens a week and charged him/her a certain token rate a minute. From the experiments, we found that prices cannot change very often, the first few minutes of a call are very important to users, and users can easily be enticed to accept and choose congestion pricing. We then used simulations to evaluate congestion pricing when there are many users using the service under the same token constraint. We used the data from the user experiments to model when a user would make calls and how he/she would react to price changes. From the simulations, we determined how we should set the congestion pricing parameters for managing thousands of users. Using the appropriate parameters, we estimate that congestion pricing can reduce call blocking rate by 60% or save provisioning by 20% while causing users to experience a price change in 4% of their usages. We further verified the user model by having real users using the service along with simulated users. We found that the real users' reactions to price changes are similar to our model for the simulated users. This methodology allows us to scale the results from a small-scale user study to evaluate a system meant to involve many more users in resource allocation decisions.

Our results for voice calls are limited because we used a token system as a budget constraint instead of charging real money. However, we believe that congestion pricing would be even more effective if real money is charged because more people would terminate their calls after a price increase. Our results are also limited in that our users are students instead of the general public.

Nonetheless, our study indicates that congestion pricing can be very effective for voice calls because of users' calling pattern and reactions to price changes. The calling pattern is very bursty, periods of high utilization and low utilization, and dominated by short duration calls. Thus congestion pricing would only need to be applied during high utilization and to long duration calls. From our user experiments, if prices do not change rapidly nor in the first few minutes of a call, then a price increase can easily entice users to terminate their calls earlier. From our simulations, if prices change slowly, it would only slightly reduce the effectiveness of congestion pricing. Thus operators can effectively use congestion pricing for voice calls because they can gain much in terms of reduction in call blocking rate or saving in provisioning while only need to have users experience price changes occasionally.

[13] Data contains 9,299 minutes of calls under flat-rate pricing and 14,110 minutes of calls under congestion pricing. The data from the congestion pricing is scaled to the data of the flat-rate pricing for comparison.

REFERENCES

[1] MacKie-Mason, J.K., and H. R. Varian. 1993. "Pricing the Internet." JFK School for Government (May).

[2] MacKie-Mason, J.K.; L. Murphy; J. Murphy. 1995. **"**The role of responsive pricing in the Internet." University of Michigan Tech. Report (June).

[3] Fishburn, P.C. and A.M. Odlyzko. 2000. "Dynamic behavior of differential pricing and quality of service options for the Internet." Decision Support Systems, 28, Elsevier, (March): 123-36.

[4] Cocchi, R.; D. Estrin; S. Shenker; L. Zhang. 1991. "A study of priority pricing in multiple service class networks." Computer Communication Review, 21, (Sept.): 123-30.

[5] Peha, J.M. 1997. "Dynamic pricing as congestion control in ATM networks." IEEE Global Telecommunications Conference, GLOBECOM '97, (New York, NY, USA). IEEE, 3: 1367-72.

[6] Murphy, L.; J. Murphy; J. K. MacKie-Mason. 1996. "Feedback and efficiency in ATM networks." IEEE International Conference on Communications. Converging Technologies for Tomorrow's Applications, ICC '96, (New York, NY, USA). IEEE, 2: 1045-9.

[7] Wang, X. and H. Schulzrinne. 2000. "An integrated resource negotiation, pricing, and QoS adaptation framework for multimedia applications." IEEE Journal on Selected Areas in Communications, 18, no. 12, IEEE, (Dec.): 2514-29.

[8] Paschalidis, I.Ch. and J. N. Tsitsiklis. 2000. "Congestion-dependent pricing of network services." IEEE/ACM Transactions on Networking, 8, no.2, IEEE/ ACM, (April): 171-84.

[9] Klausz, F.J.; D. C. Croson; R.T.A. Croson. 1998. "An experimental auction to allocate congested IT resources: the case of the University of Pennsylvania modem pool." Proceedings of the Thirty-First Hawaii International Conference on System Sciences, (Los Alamitos, CA, USA). IEEE Comput. Soc, 6: 363-73.

[10] Edell, R. and P. Varaiya. 1999. "Providing Internet access: what we learn from INDEX." IEEE Network, 13, (no.5), IEEE, (Sept.-Oct.): 18-25.

[11] ITU-T Recommendation H.323 Version 3. 1988. "Packet based multimedia communication system."

[12] Shih, J.S.; R. H. Katz; A. D. Joseph. 2001. "Pricing experiments for a computer-telephony-service usage allocation." IEEE Global Telecommunications Conference, GLOBECOM '01, (San Antonio, TX, USA). IEEE, (November).

BIOGRAPHY

Jimmy S. Shih received his B.S. and M.E. degrees from the Massachusetts Institute of Technology. He is currently a PhD candidate in Electrical Engineering and Computer Science at the University of California, Berkeley. His thesis is on applying congestion pricing to voice and data traffic.

Randy H. Katz received his undergraduate degree from Cornell University, and his M.S. and Ph.D. degrees from the University of California, Berkeley. He joined the faculty at Berkeley in 1983, where he is now the United Microelectronics Corporation Distinguished Professor in Electrical Engineering and Computer Science. His current research interests are Internet Services Architecture, Mobile Internet, and the technologies underlying the convergence of telecommunications and packet networks. Prior research interests have included: database management, VLSI CAD, and high performance multiprocessor and storage architectures.