From Self-Regulate to Admission Control in Real-Time Traffic Environment

Hengky Susanto
Department of Computer Science,
University of Massachusetts at Lowell
hsusanto@cs.uml.edu

Abstract — Network Utility Maximization (NUM) has been studied extensively to improve network quality of service (QoS) and to overcome network congestion. However, even with congestion resolved, network is not always able to meet user's minimum demand for QoS. To address this issue, this paper proposes a new "self-regulate" model that considers user tolerance for poor QoS and the *price user is willing to pay*. In addition, we demonstrate that user "self- regulate" model may not lead to convergence as today's Internet user landscape changes constantly and quickly. This leads to our proposal for an admission control algorithm which considers network revenue, user tolerance, and user utility.

Index Terms — Congestion control, network utility optimization, real time network.

I. INTRODUCTION

As Internet usage and demand for real-time traffic, such as multimedia streaming, continues to increase, network may become overwhelmed. This results in some users receiving bandwidth allocation below their minimum requirement and therefore, experience poor network quality. This, in turn, leads to low user satisfaction. To prevent this situation, Lee et al. propose in [2] that some users may *self-regulate,* i.e. abandon the network and release bandwidth when the condition is not desirable, such as when the price for bandwidth gets too high and/or when the QoS falls below the minimum. This and other previously developed self-regulate models [2,12] assume that users have zero tolerance toward poor network performance and immediately discontinue transmitting data when the QoS falls below the minimum. In [12], the authors also propose a self-regulate algorithm based on user reaction toward price fluctuation, where user immediately abandons the network when the price becomes too high.

In this paper, we propose a new user self-regulate model that is different from these earlier concepts by considering two important, and related factors: user tolerance for low quality connection and the *price per unit that user is willing to pay*. Our proposal to incorporate user tolerance for poor performance is supported by the findings of Krishnan and Sitaraman [11]. Their study shows that generally users demonstrate some tolerance toward poor network connection, and that user tolerance for poor QoS varies with the type and duration of use. Informed by this study, our self-regulate model considers a period of waiting time during which user observes whether network condition improves within a certain window of time. Importantly, we assert that user expectation of QoS and therefore, their tolerance of poor performance is related to the price a user is willing to pay. Various studies [21-25] provide the evidence that price may influence user expectation toward quality of product or service and hence, their relative tolerance of lower quality. In [23], the author observes users may be willing to adjust their expectation toward product or service quality given a lower price.

Our self-regulate model considers these user behavior insights where a user may not immediately abandon the network when QoS falls below the threshold. Rather, a user may be willing to linger and wait for some time to observe whether performance eventually improves, and his/her degree of tolerance is affected by the price he/she is willing to pay at that point of time. We investigate how these two factors may influence bandwidth allocation in order to address traffic congestion.

There are other self-regulate models, however, most do not consider closely user behavior. For instance, the authors of [16] propose self-regulation algorithm for wireless stations that maximizes the targeted overall network utilization. Authors of [19,20,27] propose self-regulation algorithm for wireless sensor nodes to maximize network utility or overcome congestion by adjusting the transmission rate according to the changes in the environment, and does not necessarily inform us on user's response to poor QoS. In [17], while the author analyses the behavior of users who self-regulate, the intent is to achieve optimal social welfare by requiring coordination among users to avoid over contributing data in peer-to-peer file sharing environment. Different from this perspective, our assumption is that users are more interested in their personal gain rather than social welfare. The authors of [18] suggest self-regulate with learning algorithm, which requires collecting data sample of network traffic from the network. However, realistically, it is unlikely that users will spend time collecting data in order to decide whether to continue the network service or otherwise.

Although our proposed self-regulate model includes user tolerance and the price user is willing to pay, it is still insufficient to bring about rate allocation convergence because user landscape in the network and network configuration change dynamically and rapidly. To address this, we propose for an admission control scheme to prevent poor QoS to resolve congestion.

We begin our proposal with problem formulation in section II. Following this, we present our major contributions: the proposal for new *self-regulate* model and admission control algorithm in section III and IV respectively. The simulation results are presented and discussed in section V, followed by concluding remarks.

II. Problem Formulation

Consider a network with a set of links L, and a set of link capacities C over the links. Given a utility function $U_s(x_s)$ of user s with the allocated bandwidth of x_s, the NUM formulation becomes

$$maximize \sum_{s\in S} U_s(x_s), \qquad (P)$$

$$s.t.\ \ Ax \le C \text{ and } x \ge \bar{0}$$

where S and A denote sets of users and routing paths, respectively, and $\bar{0}$ is a vector of zeros. A route r consists of a series of links l such that $A_{lr} = 1$ if $l \in r$ and $A_{lr} = 0$, otherwise. The sigmoidal like user utility function $U_s(x_s)$ is defined as follows.

$$U_s(x_s) = \frac{1}{(1 + e^{-\dot{a}\, x_s})}, \qquad (1)$$

where positive constant $\dot{a}$ is used as normalization. The user utility function is generally used to measure user satisfaction over bandwidth allocation, QoS, *etc*. [5,14].

The NUM formulation is solved by the Lagrangian method. However, When the utility function $U_s(x_s)$ is non-concave as in a sigmoid function, modeling a utility of a real-time, non-elastic traffic, the problem formulation becomes a non-convex problem and difficult to solve as the global maximum cannot be guaranteed [2,5,26]. Typically, a dual problem to the primal problem of (P) is constructed as follows.

$$L(x,\lambda) = \sum_{s\in S} U_s(x_s) - \lambda^T(C - Ax),$$

$$= \sum_{s\in S} U_s(x_s) - \sum_{s\in S} \lambda_s x_s + \sum_{l\in L} \lambda_l\, C_l\,,$$

where $L(x,\lambda)$ is the Lagrangian form and λ is known as a set of Lagrangian multipliers λ_l, which is often interpreted as the link cost [1,2] and

$$\lambda_s = \sum_{l\epsilon r_s} \lambda_l\,.$$

The dual problem of (P) is defined as follows.

$$minimize\ D(\lambda) \qquad (D)$$

$$s.t\ \ \lambda \ge \bar{0},$$

where the dual function

$$D(\lambda) = \max_{\bar{0}\le x\le x^{max}} L(x,\lambda)$$

and is a convex function. User decides the transmission rate $x_s(\lambda_s)$ at price λ_s by solving

$$x_s(\lambda_s) = \arg\max_{0\,\le\, x\,\le\, x^{max}} \big(U_s(x_s)\big), \qquad (2)$$

where $x_s(\lambda_s)$ denotes bandwidth allocation at price λ_s. Still, a non-concave utility function is not differentiable everywhere. A subgradient projection method is used in [2]. Thus, the network on each link l updates λ_s on that link:

$$\lambda_l^{(t+1)} = \Big[\lambda_l^{(t)} - \sigma^t\Big(C - A\,x\big(\lambda_l^{(t)}\big)\Big)\Big]^+, \qquad (3)$$

where $x\big(\lambda_l^{(t)}\big)$ denotes the rate allocation at price $\lambda_l^{(t)}$ on link l, $C - A\,x\big(\lambda_l^{(t)}\big)$ is a subgradient of problem (D), t denotes the iteration index for $0 \le t \le \infty$, and the parameter $\sigma_l^{(t)} > 0$ is the step size that controls the tradeoff between a convergence guarantee to a solution and the convergence speed. $\sigma_l^{(t)}$ is defined as follows.

$$\sigma_l^{(t)} \to 0, as\ t \to \infty \text{ and } \sum_{t=1}^{\infty} \sigma_l^{(t)} = \infty.$$

The feedback loop in the pair of equation (2) and (3) allows users to adjust their transmission rate according to the price, and for the network to control the amount of traffic flow by adjusting the price until $\lambda_l^{(t)}$ converges to a solution [2]. Furthermore, $\lambda^{(t)} \ge \lambda^{min}$, where the positive minimum price λ^{min} can be interpreted as the network's operation cost [4]. Users' adjustment of their transmission rate is affected by their varying tolerance for poor performance, which depends upon their price users are willing to pay at that point in time.

III. Self-Regulation

Our self-regulation model incorporates this concept of user tolerance and price user is willing to pay, an idea supported by various studies [11,21,23,28]. Data of network traffic collected by authors of [11] show that users may not immediately abandon the network when they experience poor connection. The same report observes that users generally are willing to *tolerate* a start-up delay for a short video streaming service for 4 seconds. Less patient users are only willing to wait for 2 seconds. They also discover that users demonstrate higher tolerance of longer start-up delay when watching long videos, such as movies. Additionally, they report that users are willing to tolerate interruptions when watching video streaming if the total interruption or "freeze" time is less than 1% of the entire video duration.

Other studies [21,23,28] indicate there is a strong correlation between user satisfaction, the price user is willing to pay, and QoS. The study in [21] reports that the price that a user is willing to pay is an indication of the extent the user is willing to tolerate poor QoS, as well as their brand loyalty. The same study shows users who are satisfied with the QoS are more willing to tolerate increasing price up to a certain level. In addition, these higher paying users are more sensitive with changes in QoS. Studies [23,28] discover users who are experiencing good QoS at the service fee within their budget

are more willing to tolerate poor QoS to a certain level within a small window of time.

To begin, we specify two assumptions underlying our proposed model of self-regulate: Users are naturally selfish and their objective is to maximize their own utility,

$$maximize\ U_s(x_s(\lambda))$$

$$over\ \ \bar{0} \le x_s^{min} \le x_s(\lambda),$$

where, x_s^{min} denotes the minimum bandwidth required by user *s*, and user solves the maximization problem above by solving (2). However, when the aggregated traffic flow exceeds the capacity, network may not be able to provide the minimum required bandwidth. Consequently, user may receive less than the minimum, such that $x_s(\lambda) < x_s^{min}$, and user utility may drop after experiencing poor network quality for some time, and eventually leave the network if the condition does not improve as reported in [11].

***Definition 1*:** User's *reasonable maximum utility* is highest degree of satisfaction achieved at or within the price the user is willing to pay.

Let m_s be the amount of money user s is willingness to spend, allowing user s to influence the amount of bandwidth allocated to him/her. Then, a "reasonable" condition of *user's affordability* is when the cost is within the range $m_s - \epsilon \le (x_s^*\ \hat{\lambda}_s) \le m_s + \epsilon$ or exactly matches user's budget. Here, x_s^* denotes the bandwidth demand for user s for $x_s^* \ge x_s^{min}$, ϵ is positive constant, $\hat{\lambda}_s$ denotes *the price per unit of utility a user is willing to pay*, and $(x_s^*\ \lambda_s)$ can be interpreted as the cost that user must incur for the service. Since user s knows the minimum bandwidth required to have a satisfying experience and his/her budget for x_s^*, user s can compute the price that he/she is willing to pay $\hat{\lambda}_s$ by solving

$$\hat{\lambda}_s = \frac{m_s}{x_s^*}.$$

In this context, m_s can be also interpreted as the budget that user s is willing to spend for x_s^*. Ideally, a user reaches *reasonable maximum utility* when $\frac{x_s \lambda_s}{m_s} = 1$, where x_s denotes the amount bandwidth allocated by network at price λ_s. It means user's budget matches the cost that user must pay for the service.

With the assumption that user is willing to wait for some time for the quality to improve and not immediately abandon the network when the quality drops below x_s^{min}, we extend our utility function $U_s(x_s(\lambda))$ to the following.

$$U_s(x_s(\lambda)) = \begin{cases} U_\delta(x_s(\lambda), \lambda, \hat{\lambda}, t), & x_s^{min} > x_s(\lambda) \\ U_s(x_s(\lambda)), & Otherwise \end{cases},$$

where $U_\delta(x_s(\lambda), \lambda, \hat{\lambda}, t)$ is the utility function which captures user utility when user experiences poor QoS over a period of time and the function $U_\delta(x_s(\lambda), \lambda, \hat{\lambda}, t)$ is defined as follows.

$$U_\delta(x_s(\lambda), \lambda, \hat{\lambda}, t) = a\, U(x_s(\lambda)) - b\, \delta_{U_s}(\lambda, \hat{\lambda}, t),$$

where a and b are constant positive variables used for normalization. Function $\delta_{U_s}(\lambda, \hat{\lambda}, t)$ can be interpreted as user dissatisfaction over duration of time t. In other words, user may become impatient when $x_s(\lambda) < x_s^{min}$, and user utility decreases over t if the quality does not improve and user continues to experience poor QoS. Function $\delta_{U_s}(, \lambda, \hat{\lambda}, t)$ is defined as follows.

$$\delta_{U_s}(\lambda, \hat{\lambda}, t) = \left(\frac{t\,(\hat{\lambda} + a'\,\lambda^{(t)})}{\lambda^{(t)}} \right)^{\beta}, \text{for } t \to \infty,$$

where the parameter $\beta \ge 0$ specifying user tolerance for the poor QoS, a' is the weight, $0 \le a' \le 1$, and $\lambda^{(t)}$ denotes λ at t. Large β means a user may be impatient; smaller β means a user is more willing to tolerate poor QoS longer. Also, a higher t value means the longer a user must wait for QoS to improve. Furthermore, the rationale for price ratio $\frac{\hat{\lambda}+\lambda}{\lambda}$ is that user tolerance for poor QoS has a reverse relationship with the price a user is willing to pay, relative to the price λ network sets. In other words, lower paying users generally have greater tolerance for poor QoS and higher paying users demonstrate lower tolerance. Authors of [28] explain this is because higher paying users have more economic latitude to switch to different service providers, while lower paying users may not. We summarize the property of patience function as follows.

The property of patience function:

A. $\delta_{U_s}(\lambda, \hat{\lambda}, t)$ is linear and decreasing, for $t \to \infty$.
B. $\delta_{U_s}(\lambda, \hat{\lambda}, t) \to -\infty$, for $t \to \infty$, and $0 \le t$.
C. $U_s(x_s(\lambda))$ is twice differentiable.
D. User with higher paying tends to be more impatient.

Chiang et al. proposes a similar function called "waiting function" in [13], where the function decreases as the waiting gets longer. He observes that estimating the parameters for waiting function of different applications is very difficult because there are too many of them. There exists a threshold δ_s^{thrd}, when excessive congestion occurs and the poor performance becomes intolerable, such that $U_s(x_s(\lambda)) < \delta_s^{thrd}$, and user self regulates to release the bandwidth and *voluntarily discharge*.

***Definition 2*:** *Voluntarily discharge* is when a user stops transmitting data when he/she is no longer can tolerate the poor performance.

The property of *voluntary discharge:*

1. Each user has voluntary discharge policy.
2. User stops transmitting data when $\delta_s^{thrd} > U_s(x_s(\lambda))$.
3. $\hat{\lambda} < \lambda$, where $\hat{\lambda} = \sum_{l \in s} \hat{\lambda}_l$ and $\lambda = \sum_{l \in s} \lambda_l$.
4. User is not willing to increase willingness to pay m_s to compensate λ when $\hat{\lambda}_s < \lambda$.

Remark: Property 3 implies the condition when the network price λ is too high or too expensive for the user. Let S_{vd} denote a set of user when voluntary discharge and stop

transmitting data after n^{th} iteration, for $n \leq t$, the following condition is satisfied:

$$\sum_{s \in S(l)} x_s^t > C_l \, , l \in L \, , \tag{4}$$

for $t \to \infty$, such that $U_{s'}^t\left(x_s(\lambda)\right) \leq \delta_s^t$, and $U_s^t\left(x_s(\lambda)\right) > \delta_s^t, s \in S_{vd}(l)$, $s' \in S_{vd}(l)$ and $S_{vd}(l) \subseteq S(l)$, where $S(l)$ refers to a set of users who are transmitting data over link l. Thus, equation (4) provides the condition for when the user to *voluntary discharge*. There exists n that satisfies the condition above, where there are users in $S_{vd}(l)$ who stop transmitting data by n^{th} iteration and there are users in set $S(l) - S_{vd}(l)$, who continue to transmit data after n^{th} iteration. The next proposition implies that, when link l is congested and $\delta_s^{thrd} \geq U_s^t\left(x_s(\lambda)\right)$, for $s \in S_{vd}(l)$, users in $S_{vd}(l)$ stop transmitting. Let $n_s(\delta_s^{thrd})$ denotes the n^{th} iteration when user utility $U_s\left(x_s(\lambda)\right) \leq \delta_s^{thrd}$. Let $U_s^t\left(x_s(\lambda_l^t)\right)$ denote utility function of user s with transmission rate x_s at price λ_l^t for link l at t^{th} iteration, for $x_s^{min} > x_s(\lambda_l^t)$.

Proposition 1: for any δ_s^{thrd} and $s \in S_{vd}(l)$, there exists $n_s(\delta_s^{thrd})$ such that $\delta_s^{thrd} \geq U_s^t\left(x_s(\lambda_l^t)\right)$ and $\sum_{s \in S(l) - S_{vd}(l)} x_s(\lambda_l^t) > C_l$, $l \in L$, for all $n_s(\delta_s^{thrd}) < t$.

Proof : We show that

$$U_s^t\left(x_s(\lambda_l^t)\right) \leq \lim_{n \to \infty} \sup U_s^t\left(x_s(\lambda_l^t)\right) = \delta_s^{thrd},$$

$$\text{for } s \in S_{vd}(l).$$

When $x_s^{min} > x_s(\lambda_l^t)$, for $t \to \infty$ and $s \in S_{vd}(l)$, $U_s^t\left(x_s(\lambda_l^t)\right) = U_\delta(\lambda_l^t, \hat{\lambda}, t) = a\ U\left(x_s(\lambda_l^t)\right) - b\ \delta_{U_s}(x_s(\lambda_l^t), \lambda_l^t, \hat{\lambda}, t) \to -\infty$.
Further, since $-\infty \leq \delta_s^{thrd}$, $U_s^t\left(x_s(\lambda_l^t)\right) \leq \delta_s^{thrd}$. We have shown that $U_s^t\left(x_s(\lambda_l^t)\right) \leq \lim sup_{n \to \infty} U_s^t\left(x_s(\lambda_l^t)\right) = \delta_s^{thrd}$. ■

Proposition 1 implies that user self-regulates according to *voluntary discharge* properties and stops transmitting data after $n_s(\delta_s^{thrd})^{th}$ iteration when $U_s\left(x_s(\lambda_l^t)\right) \leq \delta_s^{thrd}$, for $s \in S_{vd}(l)$. In essence, the above demonstrates user eventually stops transmitting data when the condition is not conducive. Next, we show that our self-regulate model also leads to convergence.

Proposition 2: Assuming that each user practices "*self-regulation*", there exists a constant step size σ in which our algorithm converges.

Proof. The proof of our self-regulate model is similar to Lee's model in [*2*]. The premise of the proof is that with the self-regulation algorithm, there exists an optimal solution λ^o of the dual. The author shows if the dual problem $D(\lambda^o)$ is differentiable at λ^o, $x(\lambda^t)$ converges to the optimal solution for primal problem, $x(\lambda^o)$. However, when it does not converge, there will be users who obey self-regulation and stop transmitting data until $x(\lambda^t)$ asymptotically converges to $x(\lambda^o)$. The rate allocation algorithm asymptotically converges to optimal solution with a constant step size σ. For detailed discussion of the proof, refer to [2]. ■

In the following step, we examine the effectiveness of self-regulation to voluntarily discharge during excessive network congestion. Let λ_l^t denote network price λ for link l at t^{th} iteration.

Proposition 3: *self-regulation* does not always lead to feasibility.

Proof: Let $S_{join}(l)$ be a *non-empty* set of users that join the network and use link l, where, $S_{join}(l) \neq S_{vd}(l)$ and $S_{join}(l) \nsubseteq S(l)$. Let us assume, in contrary, that self-regulation will *always* lead to feasibility, such that, after t^{th} iteration,

$$\sum_{s \in S(l)} x_s(\lambda_l^t) - \sum_{s' \in S_{vd}(l)} x_{s'}(\lambda_l^t) = C_l$$

and we have users in $S_{join}(l)$ join the network by t^{th} iteration . Now, we have

$$\sum_{s \in S(l)} x_s(\lambda_l^t) - \sum_{s' \in S_{vd}(l)} x_{s'}(\lambda_l^t) + \sum_{z \in S_{join}(l)} x_z(\lambda_l^t) = C_l + \sum_{z \in S_{join}(l)} x_z(\lambda_l^t),$$

where $\forall z,\ z \in S_{join}(l),\ x_z(\lambda_l^t) \geq x_z^{min} > 0$. Since $\sum_{z \in S_{join}(l)} x_z(\lambda_l^t) > 0$, than $C_l + \sum_{z \in S_{join}(l)} x_z(\lambda_l^t) > C_l$. Hence, we have a contradiction. ■

Earlier, proposition 2 implies that the algorithm converges under the assumption that there is no new user joining the network before the algorithm converges. However, as illustrated in proposition 3, when the arrival rate of new users exceeds the departure rate of existing users, the self-regulate to voluntarily discharge does not always lead to link feasibility. Furthermore, our self-regulate model shows that the algorithm may actually take even longer time to converge, which may result in higher likelihood of user composition changes before the algorithm converges.

Moreover, as self-regulation decision is controlled by and only known to user him/herself but not known to network, it is difficult for the network to distinguish between when user stops transmitting because user is self-regulating to voluntary discharge, or when he/she is taking a break from his/her activity (perhaps for a coffee break) and eventually resumes his/her activities later. The condition may cause the rate allocation algorithm to not converge because existing users only temporarily stop their activities. This phenomenon is similar to the condition where users join and leave the network before the completion of the algorithm. For these reasons, to preserve QoS when the excessive congestion occurs and the algorithm does not converge, the network may need to take an active role in implementing admission control.

IV. Admission Control

We propose an admission control mechanism to select a set of users that network can support in order to overcome the challenges of non-converging rate allocation algorithm and failure to meet user minimum QoS. The over-riding question

is: which users should be admitted to network? Ideally, network is occupied with highest paying users, who are already satisfied with the provided QoS. Conversely, the worst scenario is when network is occupied by very demanding lower paying users. Therefore, to prevent the latter, we propose a network admission control algorithm that considers network pricing, QoS, network revenue, the price user is willing to pay, and user utilization, while assuring link feasibility of the entire network.

We introduce a measurement factor ϑ for admission control, which takes QoS, user utility, and network revenue into consideration. The principal idea is to quantify the relationship and strike a balance among these three factors. A simple illustration of the principle: there are three users sharing a link with Capacity *C*. User 1 and 2 demands $\frac{C}{2}$ and willing to pay 5 unit currency for the service but currently only receive $\frac{C-1}{2}$. User 3 is willing to pay 8 unit currency for *C* but currently only receives $C - 2\left(\frac{C-1}{2}\right)$. Thus, admitting user 1 and 2 is more advantageous because they can be quickly satisfied and network can achieve a higher aggregate revenue of 10 currency units and user utility, than if just user 3 is admitted. Therefore, in our scheme, users with higher paying ability and at the same time can be quickly satisfied receive higher admission priority.

Other research work done in admission control focus and prioritize higher revenue for network. In [6], the authors propose a pricing model for admission control while maintaining QoS, but user utility is not considered in their scheme. Authors of [7] propose a solution that incorporates bandwidth allocation and penalty function into their admission control algorithm. User with high penalty score is eventually banned from transmitting data into network. Authors of [8] propose an algorithm that stops users from transmitting data when user utility falls below the threshold. Moreover, the network may also refuse admission to users who cause the network to lose revenue. Our work, on the other hand, seeks to balance both revenue and user satisfaction.

In our scheme, user's admission to network is decided by the relationship between current pricing, QoS, network revenue, and the price user is willing to pay measured through function $\vartheta_s\big(x_s(\lambda)\big)$.

$$\vartheta_s\big(x_s(\lambda)\big) = \frac{\hat{\lambda}_s}{\lambda} \ \frac{U_s\big(x_s(\lambda)\big)}{U_s(x_s^{min})}\ x_s(\lambda),$$

for $x_s^{min} \geq 0$ and $U_s(x_s^{min})$, λ, $\hat{\lambda}_s > 0$. The objective is to find a set of users with the highest ϑ_s value from

$$\vartheta_s = \vartheta_s\big(x_s(\lambda)\big).$$

One way to interpret the model is, a user with $\frac{\hat{\lambda}_s}{\lambda} > 1$ is willing to pay higher price than the current network price λ, and $\frac{U_s(x_s(\lambda))}{U_s(x_s^{min})} > 1$ can be interpreted as user is satisfied with the bandwidth allocation x_s at price λ. Hence, the admission control is answered by solving the maximization problem, which is defined as follows.

$$\max \sum_{s \in S} \vartheta_s \ z_s \tag{5}$$

$$s.t. \quad Ax\, z \leq C \ ,$$

$$z_s \in \{0,1\} \quad \forall s \in S$$

$$over \ \bar{0} \leq x^{min} \leq x,$$

where $z_s = 1$ if user *s* is selected, otherwise $z_s = 0$.

Selecting users for network admission is similar to solving the Knapsack decision problem that is, given a set of $|S|$ users, with each user associated with some value ϑ_s and the weight x_s^{min}. The objective is to select users that maximize $\sum_{s \in S} \vartheta_s$, while obeying the fixed maximum total link capacity C_l, such that $\sum_{s \in S(l)} x_s \leq C_l$. However, Knapsack decision problem turns out to be another NP-hard problem [9,10]. Thus, finding the resolution is at least as difficult as the decision problem, and there is no known polynomial algorithm which can tell, given the solution, whether it is optimal. One of the most popular approaches to solve Knapsack decision problem is by *dynamic programming* (DP) technique [10], that is to explore the space of all possible solutions by carefully decomposing things into a series of sub problems. The running time of the algorithm is $O(|S(l)|\ C_l)$. However, C_l can be very large and the computation takes long time. Consequently, user landscape in the network may change before the completion of admission control process. Thus, the existing solution for dynamic programing may be too slow for this problem. Furthermore, in dynamic programing, the weight is assumed to be integer, which is not suitable for allocated bandwidth because x_s is real number.

Therefore, to solve this network admission problem, we propose a greedy based solution. It is less optimal than dynamic programing based solutions because there is no efficient greedy rule that always constructs an optimal solution [10]. However, since admission control is done in a centralized manner, greedy based algorithm may provide a shorter running time. The algorithm is designed as the following.

Algorithm 1: *User Selection Algorithm*

1. $\vartheta_s^{\max} = \max\{\vartheta_{SET(l)}\}$
2. x_s^{min} = get_bandwidth($\vartheta_s^{\max}$)
3. **if** $(x_s^{min} + \sum_{\hat{s} \in l}\ x_{\hat{s}} \leq C_l,$ for $\forall l,\ l \in route\ r_s)$ and $(s \notin l)$ **then**
4. Reserve l for user s, for $\forall l,\ l \in r_s$
5. $C_l = C_l - x_s^{min}$ for $\forall l,\ l \in r_s$
6. $\hat{S} = \hat{S} + s$
7. $\vartheta_{SET} = \vartheta_{SET} - \{\vartheta_s^{\max}\}$
8. **Repeat** from line 1 until $|\vartheta_{SET}| = 0$ *// until ϑ_{SET} is empty*

Let ϑ_{SET} denotes a set of ϑ_s that is associated with user and $\hat{S}$ denotes a set of users admitted into network. In line 1 and 2 of

User selection algorithm, given $\vartheta_s^{\max}$, x_s^{min} is retrieved. In line 3, the algorithm verifies whether the link has sufficient capacity to provide at least x_s^{min} and verifies that x_s^{min} has not been included in previous run/loop. Line 4-6, the algorithm reserves path for the admitted user and includes user in $\hat{S}$. Next, $\vartheta_s^{\max}$ is removed from the set in line 7. However, it is also possible that some existing users may leave or incoming users may try to access the network before the network completes its own admission control. Thus, line 7 is modified as follows.

$$\vartheta_{SET} = \vartheta_{SET} - \vartheta_s^{\max} - \{\vartheta_{\text{set}}^{\text{leave}}\} + \{\vartheta_{\text{set}}^{\text{join}}\},$$

where $\{\vartheta_{\text{set}}^{\text{leave}}\}$ and $\{\vartheta_{\text{set}}^{\text{join}}\}$ are sets of users that leave and join the network prior to the completion of user selection. We also assume that network immediately provides the service as soon as the user is admitted into network.

The performance of this algorithm is determined by number of links $|L|$, number of users $|S|$, and the number of links in the path r_s of each admitted user s that needs to be updated. Thus, the total running time is at most $O(|L|.|S|)$. This selection process above is quicker than DP as we illustrate here: Assume that x_s is rounded, time required to find the entire combination of all possible solutions on link C_l is $O(|S|\ C_l)$. The time needed to find the best solution among all existing solutions is $O(|S|)$ and, for each selected s, the network must reserve the entire $l \in r_s$, which takes $O(|L|)$. in addition, there has to be another $O(|L|)$ for the network to ensure that no user is left behind. Thus, let $C^* = \max(C)$, the total running time with DP is $O\left(|L|\left((|S|.C^*) + (|S|.|L|)\right)\right)$

$$= O(|L|.|S|.C^* + |S|.|L|^2) \geq O(|L|.|S|).$$

Nevertheless, in reality, user utility is usually only known to the user but not to network. Observe that sigmoidal function $U_s(x_s(\lambda)) \to 1$ as $x_s(\lambda) \to \infty$. Thus, when $x_s^{min} < x_s(\lambda)$, $U_s(x_s^{min}) \leq 1 - \epsilon$, where ϵ denotes a positive constant variable. Notice that

$$\limsup_{x_s(\lambda)\to\infty} \frac{U_s(x_s(\lambda))}{U_s(x_s^{min}) + \epsilon} = 1.$$

Thus, we have

$$\limsup_{x_s(\lambda)\to\infty} \frac{U_s(x_s(\lambda))}{U_s(x_s^{min}) + \epsilon} \frac{\hat{\lambda}_s}{\lambda}\ x_s(\lambda) \to \frac{\hat{\lambda}_s}{\lambda}\ x_s(\lambda).$$

For that reason, function $\vartheta_s'(x_s(\lambda))$ is approximated as follows.

$$\vartheta_s'(x_s(\lambda)) \approx x_s(\lambda) \cdot \frac{\hat{\lambda}_s}{\lambda} \cdot \begin{cases} 1, & x_s(\lambda) \geq x_s^{min} \\ \dfrac{x_s(\lambda)}{x_s^{min}}, & x_s^{min} > x_s(\lambda) > 0 \\ \dfrac{\tau}{x_s^{min}} & Otherwise \end{cases}, (6)$$

where τ is a positive constant when $x_s(\lambda) = 0$. Depending on how the utility function is conditioned, user may decide not to transmit data. For instance, when the price is too high [2] or when the allocated bandwidth is less than the minimum [5].

After the completion of the admission control and if there is underutilized link, that is $\sum_{s\epsilon S(l)} x_s^{min} < C_l$, the unutilized bandwidth of $C_l - \sum_{s\epsilon S(l)} x_s^{min}$ can be distributed to admitted users by solving (5). This way, network can sell the unused bandwidth to increase revenue and throughput.

In addition to allocating unused bandwidth, the network can also adjust the bandwidth allocation that is allocated to the admitted users according to certain criteria of fairness. For example, one of the well-known fairness models is the *proportionally fairness* introduced by Kelly in [1,3].

Definition 3: A vector of rates $x = (x_l, l \in L)$ is *proportionally fair* if it is feasible, that is $x \geq 0$ and $Ax \leq C$, and if for any other feasible vector x^x, the aggregate of proportional changes is zero or negative:

$$\sum_{s\,\epsilon\, S} \frac{(x_s^* - x_s)}{x_s} \leq 0. \quad (7)$$

However, (7) is not sufficient because the network must meet the condition that $x_s^{min} \leq x^*$. Hence,

$$x^* = \begin{cases} x^*, & if\ x_s^{min} \leq x^* \\ x, & Otherwise \end{cases} \quad (8)$$

Lemma 1: Condition (8) satisfies (7).

Proof: There is other feasible vector x^* that is *proportionally fair* and $x_s^* \leq x_s^{min}$. But by condition (8), that $x_s^* = x_s$, which implies $\sum_{s\,\epsilon\, S} \frac{(x_s^* - x_s)}{x_s} = \sum_{s\,\epsilon\, S} \frac{(x_s - x_s)}{x_s} = 0$. Otherwise $\sum_{s\,\epsilon\, S} \frac{(x_s^* - x_s)}{x_s} \leq 0$. Thus, this satisfies (7). ■

V. SIMULATION

To demonstrate how the rate allocation algorithm asymptotically converges after admission control is performed, we employ a network of five nodes and four links with capacity of 10 shared by four users: user 0, 1, 2, and 3, as depicted in figure 1. The objective is to simply demonstrate how admission control is practiced and to show that the algorithm also achieves convergence after admission control. The setup configuration is described in table 1 and the initial price is 2.05 unit currency per unit bandwidth.

	User 0	User 1	User 2	User 3
Willingness to pay	40	50	60	30
Initial bandwidth (BW) demand	10	10	10	10
BW minimum requirement (x_s^{min})	2	2	2	5
Initial rate Allocation (x_s)	2.41	2.67	2.87	2.05
ϑ_s	19.16	29.40	39.36	4.16

Table 1. The simulation setup and initialization.

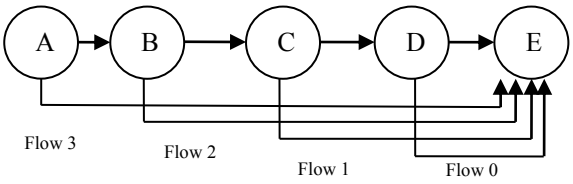

Fig. 1. Single Bottleneck.

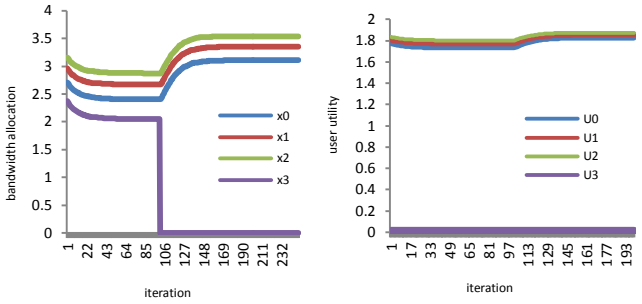

Fig. 2.a (left) and 2.b (right): Rate allocation and user utility.

As described in table 1, the total initial users' demand exceeds the link capacity of 10, that is $40 > C$. Consequently, network increases the price to 30 unit currency and force users to lower their transmission rate. Note, after the initial bandwidth allocation, user 3 suffers poor performance because network is unable to meet the required minimum bandwidth. At this time, network performs admission control by solving (6) to compute ϑ value of each user at iteration 100. The network selects a set of users according to ϑ value, resulting in the selection of user 0, 1, and 2, but not user 3 because user 3 has the least ϑ value, as described in table 1. Figure 2.a and 2.b illustrate that the rate distribution algorithm converges after user 3 is removed from the network. Next, the spare bandwidth is distributed among admitted users, user 0, 1, and 2, as illustrated in figure 2.a. Furthermore, notice that user 0, 1, and 2, receive different bandwidth allocation because the allocation is determined by the price that each user is willing to pay, which is explained in our previous work [5]. In addition, observe that user 0, 1, and 2 also achieve higher utility after user 3 is removed, as illustrated in figure 2.b. The rise in user utility reflects the increase in bandwidth allocation shown in figure 2.a. Also note that user 3 utility is flat at zero because s(he) is transmitting data below the minimum. Thus, it also makes user 3 the best candidate to be removed from the network. Once user 3 is removed, the algorithm converges as demonstrated in figure 2.b.

	Link AB	Link BC	Link CD	Link DE
Before	2.05	4.92	7.59	10
After	0	3.54	6.89	10
Adjusted	0	3.56	6.9	10

Table 2. Result of *before* and *after* implementing admission control and *adjusting* the rate according to *proportional fairness*.

We also examine the tradeoff of performing admission control by analyzing the link utilization before and after admission control is executed on each link. As illustrated in figure 2.b, user utility of each user increases after admission control is performed; however, as seen in table 2, the overall network utilization decreases. For example, as user 3 stops transmitting data after the execution of admission control, and the spare bandwidth from user 3 is distributed among the admitted users, there is zero network utilization in link AB. Similarly, network utilization also decreases in link BC and CD. Only link DE has identical throughput before and after admission control because it is the last link in the network - the one all traffic has to pass through and by which the amount of bandwidth available to be distributed is bounded. Clearly, the tradeoff of performing admission control is that network may achieve higher user utility but at the cost of maximum network utilization.

To overcome under-utilization, the network may adjust the allocated bandwidth according to proportional fairness. As shown in table 2, the network takes 0.1 from each user 0 and 1, and reallocates it to user 2 and therefore increasing user 2's flow by 0.2 according to equation (7). As a result, the link utilization on link BC and link CD increase by 0.2 and 0.1 respectively, leading to higher overall network utilization by 0.3. This affirms that Kelly's proportional fairness can be utilized to achieve higher network utilization by reallocating bandwidth among admitted users. This outcome also provides a counter example of the findings in [29,30] that proportional fairness favors shorter flows in achieving higher overall network utilization. In our simulation setup, prioritizing longer flow achieves higher overall network utilization. Thus, how network is set up determines whether proportional fairness will favor shorter or longer flows.

VI. Concluding Remarks

This paper provides new perspectives to the study of network congestion by incorporating user expectation and behavior toward quality of experience (QoE). This is important because the aggregate QoE eventually determines the level and management of the network traffic. We design a new self-regulate model that critically incorporates user tolerance for poor performance and the price user is willing to pay, closely reflecting realistic user behavior. Doing this provides us with a better understanding of how users may react and respond toward quality of connection, enabling network to make better decisions for managing network traffic. This enhanced self-regulate model, nevertheless may not lead to convergence when user landscape changes before the completion of the algorithm. This result affirms that network cannot rely solely on users to self-regulate to overcome traffic congestion.

To achieve convergence and higher user utility, we propose admission control algorithm that selects users by considering user tolerance and the price user is willing to pay. In the process of admitting users, network proactively tries to strike a balance between these criteria: QoS, user utility, and network revenue. However, finding the optimal solution and the ideal balance among these three factors is very difficult because the solution is greedy based approach. Hence, the tradeoff for a quicker solution is to settle for a suboptimal one. In addition,

we demonstrate that implementing admission control may lead to higher user utility, but it may also lead to lower network utilization or throughput. We also show that the rate allocation algorithm converges after admission control, although there is no study yet to evaluate the benefits or costs of suboptimal solution other than those mentioned in paper. These will be addressed in the future work.

ACKNOWLEDGMENT

Thank you to Prof. Byung Guk Kim for the constructive discussions, guidance, and support in writing this paper.

REFERENCES

[1] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications,* vol. 8, pp. 33-37, 1997.

[2] J. W. Lee, R.Mazumdar, and N. B. Shroff, "Non-convex optimization and rate control for multi-class services in the Internet," *IEEE / ACM Transaction on Network*, vol. 13, no. 4, pp. 827-840, August 2005.

[3] F. P. Kelly, A. Maullo, D. Tan "Rate control in communication networks: shadow prices, proportional fairness and stability", *Journal of the Operational Research Society*. 1998.

[4] A. Couch, N. Wu, and H. Susanto, "Toward a cost model for system administration," *USENIX LISA*, San Diago, 2005.

[5] H. Susanto and B. G. Kim, "Congestion Control with QoS for Real-Time Traffic", *IEEE ICCCN*, September 2013.

[6] Henrik Lundqvist, "Admission Control with Resource Reallocation," in *International Teletraffic Congress*, 2009.

[7] C. Xia, D. Towsley, and C. Zhang, "Distributed Resource Management and Admission Control of Stream Processing Systems with Max Utility," *ICDCS*, 2007, p. 68.

[8] M. Baglietto, R. Bollaa, F. Davolia, M. Marcheseb, and M. Mongellib, "A proposal of new price-based Call Admission Control rules for Guaranteed Performance services multiplexed with Best Effort traffic," *Computer Communications*, vol. 26, pp. 1470–1483, 2003.

[9] D. Williamson and D. Shmoys, *The Design of Approximation Algorithms*. New York, US: Cambridge University Press, 2011.

[10] Jon Kleinberg and Éva Tardos, *Algorithm Design*, 1st ed.: Addison-Wesley, 2005.

[11] S. Krishnan and R. Sitaraman, "Video Stream Quality Impacts Viewer Behavior: Inferring Causality using Quasi-Experimental Designs,", *The ACM Internet Measurement Conf.*, Boston, 2012.

[12] Tianshu Li, Youssef Iraqi, and Raouf Boutaba, "Pricing and admission control for QoS-enabled Internet," *Journal of Computer Network, Special issue on Internet Economics*, 2004.

[13] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "TUBE: Time Dependent Pricing for Mobile Data", *ACM SIGCOMM*, Helsinki, 2012.

[14] H. Susanto and Byung Guk Kim, "Congestion Control and User Utility Function for Real-Time Traffic", *IEEE GlobeCom-MEMS* , 2013.

[15] F. P. Kelly, A. Maullo, D. Tan "Rate control in communication networks: shadow prices, proportional fairness and stability", *Journal of the Operational Research Society*. 1998.

[16] S. Ci, M Guizani, M. Sharif, "Self-regulating network utilization in mobile ad-hoc wireless LANs", *IEEE QoS in Heterogeneous Wired/Wireless*, 2005.

[17] M. Klaas, "Over Contribution in Discretionary Databases", Workshop in the Economics of Peer-to-Peer Systems, 2004.

[18] S. Suresh, R. Savitha, and N. Sundararajan, " A Sequential Learning Algorithm for Complex-Valued Self-Regulating Resource Allocation-CSRAN", *IEEE Trans. On Neural Network. Vol 22, No 7*, July, 2011.

[19] H. Nama, M. Chiang, and N. Mandayam," Optimality Utility-Lifetime Trade-off in Self-Regulating Wireless Sensor Networks: A distributed Approach*", In proceeding of Information Science and Systems*, 2006.

[20] H. Nama, M. Chiang, and N. Mandayam,"Utility-Lifetime Trade-off in Self-regulating Wireless Sensor Networks: A Cross-Layer Design approach", *IEEE ICC*, 2006.

[21] G. Bitran, P. Oliveira, and A. Schilkrut,"Managing Customer Relationships Through Price and Service Quality", *Social Science Research Networks*, 2008.

[22] T. Ho and Y. Zheng,"Setting Customer Expectation in Service Delivery: An Integrated Marketing-Operations Perspective", *Journal of the Institute for Operations Research and management Science*, 2002.

[23] B. Johnston, "The zone of tolerance: Exploring the relationship between service transactions and satisfaction with the overall service", *International Journal of Service Industry Mana*gement, Vol. 6 Iss: 2, pp.46 – 61, 1995.

[24] N. Sharma, D. Drain, E. Cudney, and K. Ragsdell, "Predicting Warranty Cost on the Basis of Customer Expectation and Product Performance for Nominal the Best Characteristics", *Journal of Industrial and Systems Engineering*, Vol. 2 (2), pp. 97-113, 2008.

[25] B. Kaushik, H. Zhang, X. Fu, B. Liu, and J. Wang, "Smartparcel: A collaborative data sharing framework for mobile operating systems," in *Distributed Computing Systems Workshops (ICDCSW)*, 2013.

[26] S. Boyd and L. Vandenberghe, "Convex Optimization", *Cambridge University Press*, 2003.

[27] P. Levis, N. Patel, D Culler, and S. Shenker," Trickle: A Self-Regulating Algorithm for Code Propagation and Maintenance in Wireless Sensor Networks", *IEEE NSDI*, 2004.

[28] E. Anderson,"Customer Satisfaction and Price Tolerance", *Marketing Letters*, Vol 7, issue 3, July 1996.

[29] P. Gevros, J. Criwcrift, P. Kristein, and S. Bhatto, "Congestion Control Mechanisms and the Best Effort Service Model", *IEEE Journal on Network*, Vol. 15, Issue 3, 2001.

[30] K. Magoutis, P. Sarkar, G. Shah, "OASIS: Self-tuning storage for application", *IEEE/NASA Conference on Mass Storage Systems and Technologies*. 2006.