상이한 네트워크 서비스 어떻게 향상시킬까?

김용재* · †이석준* · 임재익**

How to Reinvent Network Services for All

Yong J. Kim* · †Seojun Lee* · Jay Ick Lim**

■ Abstract ■

Besieged by needs for upgrading the current Internet, social pressures, and regulatory concerns, a network operator may be left with few options to improve his services. Yet he can still consider a transition prioritizing network services. In this paper, we describe a transition from a non-priority system to a prioritized one, using non-preemptive M/G/1 model. After reviewing the constraints and theoretical results from past research, we describe steps making the transition Pareto-improving, which boils down to a multi-goal search for a Pareto-improving state. We use a genetic algorithm that captures actual transition costs along with incentive-compatible and Pareto-improving constraints. Simulation results demonstrate that the initial post-transition solutions are typically Pareto-improving. For non Pareto-improving solutions, the heuristic quickly generates Pareto-improving and incentive-compatible solutions.

Keywords : Network Management, Priority Queue, Differential Treatment of Services, Genetic Algorithm

1. Introduction

In recent years, the "net neutrality" has surfaced only to rekindle the interest in preferential treatments of heterogeneous Internet traffic. With applications such as IPTV taking up larger

논문접수일 : 2008년 10월 24일 논문게재확정일 : 2008년 11월 07일
* 건국대학교 경영정보학과 교수
** 아주대학교 경영대학 교수
† 교신저자

chunks of Internet traffic, a few network carriers have implicitly considered or planned differential Internet services to control the unwieldy traffic growth. However, such moves have been met with public protests and regulatory concerns both in the US and in Korea. They fret that prioritized networks would drive out less fortunate users and induce discrimination, directly suffocating the free spirit of the original Internet. Joining the camp are Internet portals and service providers like Yahoo and Google that insist that the carriers should remain neutral to the Internet traffic regardless of source, destination, traffic volume, or time of a day. Although incidents such as carriers refusing VoIP traffic were quickly noticed and subsequently resolved by regulating authorities, whether the net neutrality should be maintained or not will be recurring in the foreseeable future because the Internet is still rapidly evolving and it is hard to expect what it will take to get this matter settled (Kwak 2006; Laxton 2006; Hahn and Wallsten 2006).

Furthermore, much to the chagrin of many researchers and network carriers, upgrading the current Internet on a global basis has been found extremely difficult for lack of compatible network protocols and exorbitant costs. Complicating the situation is the fact that the market is driven by network carriers confined in their home lands, regulated by different laws and governmental agencies, and under different market conditions. As a result, despite that network carriers want to expand network capacity and to accommodate new QoS protocols—all geared to help smooth growth of the Internet—, carriers seem to be left with few options to exercise to change their status quo.

The primary goal of this paper is to illustrate that a network carrier still has a way to improve its operation without making any customers worse off, thus making itself immune from the net neutrality controversy. For example, delay-sensitive VoIP service can have a higher priority over other Internet applications like e-mail so that VoIP traffic does not suffer transmission delays. To illustrate such potentials in communication networks we use a multi-class M/G/1 queuing model approximating a queueing network owned by a network carrier who is a monopolist maximizing net system value defined as the sum of values of finished jobs minus the total network delay costs.

Combating network delay has a long history and a substantial body of research suggests that networks better be operated as prioritized (multi-class) rather than non-priority (single-class) systems. For example, empirical studies (Edell and Varaiya 1999; Dovrolis and Ramanathan 1999; Cochi 1993) demonstrated the needs and benefits of prioritized operation. In economics literature, it was Pigou (1920) who first studied the queueing delay effect in a congestible resource and Naor (1969), Knudsen (1972), and Mendelson (1985) advanced the idea in different contexts. The model used in this paper largely borrows from Mendelson and Whang (1990) who showed that priority-and time-dependent pricing induces individual users to select a correct priority class and that the resulting state is both optimal and incentive-compatible. In the area of network management, although there have been a few papers that address transition issues similar to ours (for example, see Cochi et al., 1993), none of them have examined welfare aspects of the transition from a non-priority to a prioritized system rigorously, which is the primary focus of

this paper.

The plan of this paper is as follows. First, we describe the system and state fundamental theorems showing the superiority of prioritized system in terms of total delay cost, the full price (also known as social marginal cost) faced by individual jobs, and revenue. We then show that not every transition is Pareto-improving in the sense that there are some user classes disadvantaged by the transition, which drives us to devise a theoretical model incorporating transition cost, an externality cost impacting on individual users after the transition. However, the computational complexity associated with the transaction cost motivates us to develop an efficient heuristic to search for Pareto-improving transitions. Through a simulation using a genetic algorithm, we test the quality of post-transition solutions and the quality of solutions generated by the heuristic. Discussing the characteristics of the solutions then follows before concluding remarks and future research directions.

2. Description of the System-Before and After Prioritization

In this section, we describe the system before and after the transition along with results central to our discussion. We assume that arrival of jobs to the network is governed by N independent Poisson processes, where class-i jobs arrive at rate λ_i. Following Mendelson and Whang (1990), let $V_i(\lambda_i)$ denote the contribution of class-i jobs when the class's arrival rate to the system is λ_i where $V_i(\lambda_i)$ is monotone increasing, continuously differentiable, and strictly concave. The marginal class-i user's valuation of a completed job will be $\partial V_i(\lambda_i)/\partial\lambda_i$ and the system value function, $V(\underline{\lambda})$, is defined as the sum of individual classes, i.e. $V(\underline{\lambda})=\sum_{k=1}^{N} V_k(\lambda_k)$ where $\underline{\lambda}=(\lambda_1, \cdots, \lambda_N)$.

Jobs are served on FCFS basis and class-i service time distribution is generally distributed with mean c_i and second moment $c_i^{(2)}$. If the network is run as a non-priority M/G/1, the expected sojourn time of class-i of non-priority system, ST_i^1, is $ST_i^1(\underline{\lambda})=\Lambda_N/\overline{S_N}+c_i$ where $S_i=\sum_{k=1}^{i}\lambda_k c_k$, $\overline{S_i}=1-S_i$, and $\Lambda_i=\sum_{k=1}^{i}\lambda_k c_k^{(2)}/2$ (Kleinrock 1976). If v_i is the delay cost per unit time for a class-i user, the total delay cost is then defined as $TC^1(\underline{\lambda})=\sum_{k=1}^{N} v_k\lambda_k ST_k^1(\underline{\lambda})$ and the net value maximizing problem is to choose $\underline{\lambda}=(\lambda_1, \cdots, \lambda_N)$ to maximize $\sum_{k=1}^{N}\{V_k(\lambda_k)-v_k\lambda_k(\Lambda_N/\overline{S}_N+c_k)\}$ so that the optimal arrival rate vector $\underline{\lambda}^+=(\lambda_1^+, \cdots, \lambda_N^+)$ satisfies the first-order conditions $\partial V_i(\lambda_i)/\partial\lambda_i=v_i ST_i^1(\underline{\lambda})+\sum_{k=1}^{N} v_k\lambda_k\partial ST_k^1(\underline{\lambda})/\partial\lambda_i$.

Assuming that at least one internal solution exists under the demand relation, the class-i externality cost be equated with the optimal price, i.e., $p_i^+(\underline{\lambda}^+)=\sum_{k=1}^{N} v_k\lambda_k\partial ST_k^1(\underline{\lambda}^+)/\partial\lambda_i$. As Pigou (1920) and other researchers found, $p_i^+(\underline{\lambda}^+)$ is to induce class-i users to pay the full price (i.e., the social marginal cost) but may fail to be incentive-compatible because a class-i user will select $p_k^1(\underline{\lambda}^+)$ whenever $p_k^1(\underline{\lambda}^+)<p_i^1(\underline{\lambda}^+)$ and $p_k^1(\underline{\lambda}^+)=\min\{p_1^1(\underline{\lambda}^+), \cdots, p_N^1(\underline{\lambda}^+)\}$. Consequently the system may not reach an optimal state and we need to introduce a time-dependent pricing such as $p^1(t)=(t^2/2\overline{S}_N^+ + \Lambda_N^+ t/\overline{S}_N^{+2})\sum_{k=1}^{N} v_k\lambda_k^+$ where $\overline{S}_i^+=1-\sum_{k=1}^{i}\lambda_k^+ c_k$ and $\Lambda_i^+=\sum_{k=1}^{i}\lambda_k^+ c_k^{(2)}/2$

for $i=1, \cdots, N$ (Kim and Mannino 2003).

Now suppose that the network carrier decides to transform the non-priority system to a non-preemptive priority M/G/1 and to apply the so-called v_i/c_i priority assignment rule such that $v_i/c_i \geq v_j/c_j$ if $i<j$ for $i, j \in \{1, \cdots, N\}$ and class-i users have the higher priority over class-j users. The rule is optimal for a nonpreemptive M/G/1 with a given $\underline{\lambda}=(\lambda_1, \cdots, \lambda_N)$. The expected sojourn time of class-i, ST_i is $ST_i(\underline{\lambda})=\Lambda_N/\bar{S}_i\bar{S}_{i-1}+c_i$ (Keinlock 1976) and the total delay cost, $TC(\underline{\lambda})$, is the sum of delay costs of individual classes : $TC(\underline{\lambda})=\sum_{k=1}^{N} v_k\lambda_k(\Lambda_N/\bar{S}_k\bar{S}_{k-1}+c_k)$. The net value maximizing problem for the prioritized system is to choose $\underline{\lambda}=(\lambda_1, \cdots, \lambda_N)$ to maximize $\sum_{k=1}^{N}(V_k(\lambda_k)-v_k\lambda_k ST_i(\underline{\lambda}))$. Assuming that the optimal arrival rate vector $\underline{\lambda}^{\square}=(\lambda_1^{\square}, \cdots, \lambda_N^{\square})$ for the first-order conditions

$$\partial V(\underline{\lambda})/\partial\lambda_i = v_i ST_i(\underline{\lambda}) + \left\{\sum_{k=1}^{N}\frac{v_k\lambda_k c_i^{(2)}}{2\bar{S}_{k-1}\bar{S}_k}+\frac{v_i\lambda_i c_i\Lambda_N}{\bar{S}_{i-1}\bar{S}_i^2}+\sum_{k=i+1}^{N}\left(\frac{v_k\lambda_k c_i\Lambda_N}{\bar{S}_{k-1}^2\bar{S}_k}+\frac{v_k\lambda_k c_i\Lambda_N}{\bar{S}_{k-1}\bar{S}_k^2}\right)\right\}$$

exists, the optimal price for class-i, $p_i(\underline{\lambda}^{\square})$, will be set as

$$p_i(\underline{\lambda}^{\square}) = c_i\left(\frac{v_i\lambda_i^{\square}\Lambda_N^{\square}}{\bar{S}_{i-1}^{\square}\bar{S}_i^{\square 2}}+\sum_{k=i+1}^{N}\left(\frac{\Lambda_N^{\square}v_k\lambda_N^{\square}}{\bar{S}_k^{\square}\bar{S}_{k-1}^{\square 2}}+\frac{\Lambda_N^{\square}v_k\lambda_k^{\square}}{\bar{S}_{k-1}^{\square}\bar{S}_k^{\square 2}}\right)\right) + \frac{c_i^{(2)}}{2}\left(\sum_{k=1}^{N}\frac{v_k\lambda_k^{\square}}{\bar{S}_{k-1}^{\square}\bar{S}_k^{\square 2}}\right)$$ where $\bar{S}_i^{\square}=1-\sum_{k=1}^{i}\lambda_k^{\square}c_k$

and $\Lambda_i^{\square}=\sum_{k=1}^{i}\lambda_k^{\square}c_k^{\square}/2$ respectively. Yet the above optimal prices are not incentive-compatible (Mendelson and Whang 1990), and a priority-and time-dependent pricing scheme should be used :

$$p_i(t)=t\left(\frac{v_i\lambda_i^{\square}\Lambda_N^{\square}}{\bar{S}_{i-1}^{\square}\bar{S}_i^{\square 2}}+\sum_{k=i+1}^{N}\left(\frac{\Lambda_N^{\square}v_k\lambda_k^{\square}}{\bar{S}_k^{\square}\bar{S}_{k-1}^{\square 2}}+\frac{\Lambda_N^{\square}v_k\lambda_k^{\square}}{\bar{S}_{k-1}^{\square}\bar{S}_k^{\square 2}}\right)\right) + \frac{t^2}{2}\left(\sum_{k=1}^{N}\frac{v_k\lambda_k^{\square}}{\bar{S}_{k-1}^{\square}\bar{S}_k^{\square}}\right)$$ (Kim and Mannino 2003).

In the subsequent discussion, we assume that time-dependent $p_i^1(t)$ and $p_i(t)$ are always employed to focus on transitional issues around the two equilibria $\underline{\lambda}^{+}$ and $\underline{\lambda}^{\square}$.

3. Fundamental Transition Theorems

This section summarizes a few theoretical results supporting prioritization of network services. The first theorem rephrases what Mitrani (1998) found : if the network moves from a non-priority M/G/1 to a nonpreemptive M/G/1, the total delay cost of the nonpreemptive M/G/1, given the fixed arrival rate vector, will be less than that of the non-priority.

Theorem 1 (Delay Costs)

Given a fixed arrival rate vector $(\lambda_1, \cdots, \lambda_N)$, *the transition from* non-priority M/G/1 *to nonpreemptive priority* M/G/1 *results in a lower total delay cost. In other words,*

$$\sum_{k=1}^{N} v_k\lambda_k ST_i(\underline{\lambda}) < \sum_{k=1}^{N} v_k\lambda_k ST_k^1(\underline{\lambda}).$$

The theorem indicates that the net system value will increase after the transition because

$$V(\underline{\lambda}^{+})-\sum_{k=1}^{N} v_k\lambda_k ST_k^1(\underline{\lambda}^{+})$$

$$< V(\underline{\lambda}^{+})-\sum_{k=1}^{N} v_k\lambda_k ST_k(\underline{\lambda}^{+})$$

$$\leq \max_{\underline{\lambda}}\left\{V(\underline{\lambda})-\sum_{k=1}^{N} v_k\lambda_k ST_k(\underline{\lambda})\right\}$$

$$\leq V(\underline{\lambda}^{\square})-\sum_{k=1}^{N} v_k\lambda_k^{\square} ST_k(\underline{\lambda}^{\square})$$; the first '<' is due to the theorem, the second '≤' by definition, and the

third '≤' due to the fact that $\underline{\lambda}^{\square}$ solves the net value maximizing problem. Therefore the following corollary holds.

Corollary 1 : (Increased Net System Value)

The transition from nonpreemptive non-priority M/G/1 *to nonpreemptive* M/G/1 *will result in the increased net system value.*

Note that the individual users are assumed to be rational and they make their decisions based on the price and delay cost experienced in the network. The key question is whether the sum of price and delay cost (viz. full price or so-called social marginal cost) will increase after the transition because net value maximizing may make some of them worse off. In the context of network neutrality, would any of Internet services be negatively affected after the transition? The next theorem, whose proof is too tedious and omitted here (but can be secured by asking authors), tries to answer the question qualitatively.

Theorem 2 : (Full Prices)

For a fixed traffic $\underline{\lambda}$*, the sum of price and sojourn time cost of nonpreemptive* M/G/1 *is less than that of non-priority* M/G/1*. In other words,* $\partial TC^{1}(\underline{\lambda})/\partial\lambda_i < \partial TC(\underline{\lambda})/\partial\lambda_i$ *for* $i=1, \cdots, N$.

Theorem 2 simply says, by turning to the prioritized system, the network operator can decrease the full price of individual users as long as the arrival rate vector remains unchanged. However we cannot assume that the arrival rate vector does so after the transition as the following example illustrates.

Example 1 :

Consider a system that has two classes with $\partial V_1(\lambda_1)/\partial\lambda_1 = 9-20\lambda_1$ on $\lambda_1 \in [0,\ 0.45]$, and $\partial V_2(\lambda_2)/\partial\lambda_2 = 12-30\lambda_2$ on $\lambda_1 \in [0,\ 0.4]$. Let $v_1=2$, $v_2=1$, $c_1=0.1$, $c_2=2$, $c_1^{(2)}=2c_1^2$, and $c_2^{(2)}=2c_2^2$ respectively, indicating that the class-1 users are more sensitive to delays. By solving the two first order conditions

$$\partial V_i(\lambda_i)/\partial\lambda_i = v_i ST_i^1(\underline{\lambda}) + \sum_{k=1}^{N} v_k\lambda_k \partial ST_i^1(\underline{\lambda})/\partial\lambda_i \quad \text{and}$$

$$\partial V(\underline{\lambda})/\partial\lambda_i = v_i ST_i(\underline{\lambda}) + \left\{ \begin{array}{l} \sum_{k=1}^{N} \dfrac{v_k\lambda_k c_i^{(2)}}{2\bar{S}_{k-1}\bar{S}_k} + \dfrac{v_i\lambda_i c_i \Lambda_N}{\bar{S}_{i-1}\bar{S}_i^2} \\ + \sum_{k=i+1}^{N} \left(\dfrac{v_k\lambda_k c_i \Lambda_N}{\bar{S}_{k-1}^2\bar{S}_k} + \dfrac{v_k\lambda_k c_i \Lambda_N}{\bar{S}_{k-1}\bar{S}_k^2} \right) \end{array} \right\}$$

respectively, we obtain $\underline{\lambda}^{+}$ = (0.3754, 0.1110) and $\underline{\lambda}^{\square}$ = (0.3718, 0.1517).

A serious problem with the transition in the example is that class-1 users' optimal price increases from 0.082 to 0.096 while that of class-2 users decreases from 6.06 to 4.49. Worse yet, class-1 users are worse off (their full price increases from 1.492 to 1.565) while class-2 users' full price (defined as the sum of access charge and sojourn time cost) reduces from 8.669 to 7.449. The public will notice that the transition favors class-2 users over class-1. In short, the transition is not Pareto-improving.

4. The Model for the Transition Problem

In this section, we describe a thought process for obtaining an exact solution incorporating the two externalities caused by delays and Pareto-improving constraints before we move to the ensuing approximation method. In order to deal with the transition discriminating against class-1 users, we might want to add constraints that guarantee the same or lower level of full price to class-1 users after the transition. To general-

ize the problem, suppose that class-i users are worse off after the transition. Initially, we can consider a constraint guaranteeing that class-i users will have the same or lower level of *full price* after the transition as follows :

$$v_i ST_i^1(\underline{\lambda}^+) + \sum_{k=1}^{N} v_k \lambda_k^+ \partial ST_k^1(\underline{\lambda}^+)/\partial \lambda_i \geq \partial TC(\underline{\lambda})/\partial \lambda_i$$

The problem can be rewritten as a Lagrangean problem with a multiplier vector $\underline{\theta}=(\theta_1, \cdots, \theta_N)$:

$$\max_{\underline{\lambda},\underline{\theta}} \left\{ \begin{array}{l} \sum_{j=1}^{N} \left(V_j(\lambda_j) - TC(\underline{\lambda}) \right) - \\ \sum_{j=1}^{N} \theta_j \left(\begin{array}{l} \partial TC(\underline{\lambda})/\partial \lambda_j - v_j ST_j^1(\underline{\lambda}^+) - \\ \sum_{k=1}^{N} v_k \lambda_k^+ \partial ST_k^1(\underline{\lambda}^+)/\partial \lambda_j \end{array} \right) \end{array} \right\}$$

If $\underline{\lambda}^o = \underline{\lambda}^o = (\lambda_1^o, \cdots, \lambda_N^o)$ and $\underline{\theta}^o = (\theta_1^o, \cdots, \theta_N^o)$ solve the problem, the necessary conditions are

$$V_i(\lambda_i^o) = v_i ST_i(\underline{\lambda}^o) + \sum_{k=1}^{N} v_k \lambda_k^o \partial ST_k(\underline{\lambda}^o)/\partial \lambda_i \quad \text{and}$$

$$+ \sum_{k=1}^{N} \theta_k^o \partial^2 TC(\underline{\lambda}^o)/\partial \lambda_i \partial \lambda_k$$

$$\theta_i^o \left(\begin{array}{l} \partial TC(\underline{\lambda}^o)/\partial \lambda_i - v_i ST_i^1(\underline{\lambda}^+) \\ - \sum_{k=1}^{N} v_k \lambda_k^+ \partial ST_k^1(\underline{\lambda}^+)/\partial \lambda_i \end{array} \right) = 0$$

for $i=1, \cdots, N$ respectively.

Unfortunately $(\underline{\lambda}^o, \underline{\theta}^o)$ may not lead to a Pareto-improving state. Re-solving the system described in Example 1 with the newly introduced constraints above, we obtain $(\lambda_1^o, \lambda_2^o) = (0.375, 0.144)$ and class-1 users' full price is 1.506, still greater than 1.492, the class-1 full price before the transition. This seemingly perplexing result can be explained by carefully examining the new constraints : class-i users are affected by the interactions not only through priority queues but also through the new constraints mandating status quo before the transition. The network operator needs to internalize the rippling effects by integrating the interactions into the existing delay costs.

In order to derive the system cost rigorously, let $RTC(\underline{\lambda})$ denote the *revised* total system cost for the system, incorporating the total delay cost and the Pareto-improving constraints. Conceptually, with the revised total cost $RTC(\underline{\lambda})$, the marginal total cost caused by a class-i user entering the network would be $\partial RTC(\underline{\lambda})/\partial \lambda_i$ and the transition is Pareto-improving if

$$\partial RTC(\underline{\lambda})/\partial \lambda_i \leq v_i ST_i^1(\underline{\lambda}^+) + \sum_{k=1}^{N} v_k \lambda_k^+ \partial ST_k^1(\underline{\lambda}^+)/\partial \lambda_i .$$

Combining the Pareto-improving constraints with appropriate Lagrangean multipliers, we write the revised total cost in a recursive form (i.e., differential equation) :

$$RTC(\underline{\lambda}) = \sum_{k=1}^{N} v_k \lambda_k ST_k(\underline{\lambda}) + \sum_{j=1}^{N} \theta_j \left(\begin{array}{l} \partial RTC(\underline{\lambda})/\partial \lambda_j - v_j ST_j^1(\underline{\lambda}^+) \\ - \sum_{k=1}^{N} v_k \lambda_k^+ \partial ST_k^1(\underline{\lambda}^+)/\partial \lambda_j \end{array} \right)$$

The first summation represents the total delay cost and the second the Pareto-improving constraint weighted by the Lagrangean multiplier vector $\underline{\theta}=(\theta_1, \cdots, \theta_N)$. Of course, $\underline{\theta}$ will be determined *after* we solve the net system value maximizing problem

$$\max_{\underline{\lambda},\underline{\theta}} \left\{ \sum_{k=1}^{N} V_k(\lambda_k) - RTC(\underline{\lambda}) \right\} \quad \text{where}$$

$$RTC(\underline{\lambda}) = \sum_{k=1}^{N} v_k \lambda_k ST_k(\underline{\lambda}) + \sum_{j=1}^{N} \theta_j \left(\begin{array}{l} \partial RTC(\underline{\lambda})/\partial \lambda_j - v_j ST_j^1(\underline{\lambda}^+) - \\ \sum_{k=1}^{N} v_k \lambda_k^+ \partial ST_k^1(\underline{\lambda}^+)/\partial \lambda_j \end{array} \right) \quad \text{for} \quad \lambda_i \geq 0,$$

$\theta_i \geq 0$. If $(\underline{\lambda}^*, \underline{\theta}^*)$ solves the net value max-

imizing problem, the initial condition for the differential equations for the revised total cost will be $RTC(\underline{\lambda}^{\bullet}) = \sum_{k=1}^{N} v_k \lambda_k^{\bullet} ST_k(\underline{\lambda}^{\bullet})$ owing to the Kuhn-Tucker's complimentary slackness conditions. However, the system of differential equations is intractable. As such, we propose a heuristic delivering the same or lower level of full prices across all classes; if class-i users suffer from the transition, add Δ_i to the post-transition full price, $\partial TC(\underline{\lambda})/\partial\lambda_i$, so that

$$v_i ST_i^1(\underline{\lambda}^+) + \sum_{k=1}^{N} v_k \lambda_k^+ \partial STC_k^1(\underline{\lambda}^+)/\partial\lambda_i \geq \partial TC(\underline{\lambda})/\partial\lambda_i + \Delta_i$$

and $\theta_i(v_i ST_i^1(\underline{\lambda}^+) + \sum_{k=1}^{N} v_k \lambda_k^+ \partial ST_k^1(\underline{\lambda}^+)/\partial\lambda_i - \partial TC(\underline{\lambda})/\partial\lambda_i - \Delta_i) = 0.$

The intuition behind the heuristic is simple : considering that the additional constraints are extra cost, we simply add Δ_i to the total delay cost to internalize the cost of Pareto-improving constraints. In other words, the problem is reformulated as

$$\max_{\underline{\lambda},\underline{\theta}} \begin{pmatrix} \sum_{k=1}^{N} V_k(\lambda_k) - \sum_{k=1}^{N} v_k \lambda_k ST_k(\underline{\lambda}) \\ -\sum_{j=1}^{N} \theta_j \begin{pmatrix} \partial TC(\underline{\lambda})/\partial\lambda_j + \Delta_j - v_j ST_j^1(\underline{\lambda}^+) \\ -\sum_{k=1}^{N} v_k \lambda_k^- \partial ST_k^1(\underline{\lambda}^+)/\partial\lambda_j \end{pmatrix} \end{pmatrix} \quad (1)$$

If $(\underline{\lambda}^*, \underline{\theta}^*)$ is the solution to (1), the Kuhn-Tucker necessary conditions are

$$V_i(\underline{\lambda}^*) = v_i ST_i(\underline{\lambda}^*) + \sum_{k=1}^{N} v_k \lambda_k^* \partial ST_k(\underline{\lambda}^*)/\partial\lambda_i + \sum_{j=1}^{N} \theta_k^* \partial^2 TC(\underline{\lambda}^*)/\partial\lambda_j \partial\lambda_i$$ and the supplementary conditions are

$\theta_i^*(\partial TC(\underline{\lambda}^*)/\partial\lambda_i + \Delta_i - v_i ST_i^1(\underline{\lambda}^+) - \sum_{k=1}^{N} v_k \lambda_k^+ \partial ST_k^1(\underline{\lambda}^+)/\partial\lambda_i) = 0$. And the optimal price for class-i users is

$$p_i(\underline{\lambda}^*) = \sum_{k=1}^{N} v_k \lambda_k^* \partial ST_k(\underline{\lambda}^*)/\partial\lambda_i + \sum_{k=1}^{N} \theta_k^* \partial^2 TC(\underline{\lambda}^*)/\partial\lambda_k \partial\lambda_i .$$

If $\underline{\lambda}^*$ and $\underline{\theta}^*$ are solution vectors for the necessary conditions, we check whether Pareto-improving constraints are met. If the constraints are not met, we repeatedly adjust $\underline{\Delta} = (\Delta_1, \cdots, \Delta_N)$ until the constraints are satisfied. The genetic algorithm discussed later seeks for acceptable $\underline{\Delta} = (\Delta_1, \cdots, \Delta_N)$ as illustrated in the next example.

Example 2 :

Suppose we solve the identical problem in Example 1 with the Pareto-improving constraints added. Because class-1 users get worse off after prioritization, we assume that $\theta_1 > 0$ and $\theta_2 > 0$. For different values of Δ_1, we can calculate λ_1, λ_2, and θ_1 to ensure improved net system values and lowered full prices.

One last issue is about the incentive-compati-

〈Table 1〉 Pareto-improving transitions for different Δ_1 values

Δ_1	$(\lambda_1, \lambda_2, \theta_1)$	Net System Value	Full Price	Revenue
0.08	(0.378,0.135,0.094)	2.437	(1.440,7.851)	0.735
0.09	(0.378,0.134,0.099)	2.436	(1.431,7.884)	0.735
0.10	(0.379,0.133,0.106)	2.435	(1.423,7.917)	0.735
Non Priority	(λ_1, λ_2) = *(0.375,0.111)*	*2.298*	*(1.492,8.669)*	*0.704*

bility of the solution because the optimal price also needs to be incentive-compatible. As we noted, the θ_k^* terms are numerically determined *after* solving the necessary conditions and the incentive compatibility should be checked on at that point. For that purpose, let us define the cheating penalty function $\Pi^i(j)$, representing the extra cost that a class-i user should bear when selecting class-j priority over class-i priority, as below (Mendelson and Whang 1990).

$$\Pi^i(j) = E_i[p_j^*(t)] + v_i(ST_j(\underline{\lambda}^*) - c_j + c_i) - E_i[p_j^*(t)] - v_i ST_i(\underline{\lambda}^*)$$

where $p_j^*(t) = t\left(\frac{v_i\lambda_i^*\Lambda_N^*}{S_{i-1}^*S_i^{*2}} + \sum_{k=i+1}^{N}\left(\frac{\Lambda_N^* v_k\lambda_k^*}{S_k^*S_{k-1}^{*2}} + \frac{\Lambda_N^* v_k\lambda_k^*}{S_{k-1}^*S_k^{*2}}\right)\right) + \frac{t^2}{2}\left(\sum_{k=1}^{N}\frac{v_k\lambda_k^*}{S_{k-1}^*S_k^*}\right) + \sum_{k=1}^{N}\theta_k^*\partial^2 TC(\underline{\lambda})/\partial\lambda_k\partial\lambda_i$ and

$E_i[p_j^*(t)]$ calculates the expected price that a class-i user will pay when she joins class-j.

Using the cheating penalty values, we can test the incentive-compatibility of optimal prices *after* reaching a Pareto-improving state $\underline{\lambda}^*$ (ex ante). As an illustration, <Table 2> lists $(\Pi^1(2), \Pi^2(1))$ for different Δ_1's satisfying the incentive-compatibility. Compounded by the search process with the incentive compatibility conditions, the acceptable range of Δ_1 values is expected to reduce.

5. Genetic Algorithm and Simulation Study

As we noted in the previous section, finding an acceptable $\underline{\Delta}$ can be viewed as a search problem. In this section, we compare the quality of solutions with simple random selection of Δ_1 $(i=1, \cdots, N)$ values to that of systematic selection using a genetic algorithm. Before presenting the results of a simulation study to evaluate the quality of solutions, we briefly describe the genetic algorithm.

<Table 2> Cheating Penalties for Selected Δ_1 Values

Δ_1	$(\Pi^1(2), \Pi^2(1))$	Net Value	Full Prices
0.00	(0.922, 0.510)	2.442	(1.498,7.720)
0.08	(1.281, 0.032)	2.437	(1.490,7.753)
0.09	(1.325,-0.027)	2.436	(1.486,7.785)
0.10	(1.459,-0.204)	2.435	(1.473,7.818)

5.1 Genetic Algorithm

For the underlying objective function of the problem, we decided to use a vector $\Delta_1 = (i=1, \cdots, N)$ as the problem representation where $\Delta_i = 0$ if class-i is better off and $\Delta_i > 0$ otherwise. For an N-class problem with $N'(N>N')$ worse-off classes due to prioritization, the genetic algorithm contains a string of N' base 10 genes.

We devised three fitness functions to experiment with the interaction between the system objective and constraint satisfaction. Solving constrained optimization problems can be a challenge for genetic algorithms. Generating constraint-satisfying solutions is often easier than solving a constrained optimization problem because there is no simple way to design a fitness function with two disparate measures (objective function and constraint satisfaction). Thus, we use constraint satisfaction for the first fitness function as shown as below where SMC_i^1 and SMC_i^1 represent the social marginal costs (i.e., the sum of delay cost and price) of class-i before and after the transition.

$$CS = \sum_{i=1}^{N'} PI_i/N' + \sum_{i=1}^{N}\sum_{j=1, \neq i}^{N} IC_{ij}/N(N-1)$$

where (2)

$PI_i = 1$ if class-i is better off

$= (SMC_i^1 - SMC_i^*)/SMC_i^1$ otherwise

$IC_{ij} = 1$ if $\Pi^i(j) \geq 0$

$= (|MinCP| - |\Pi^i(j)|)/|MinCP|$ otherwise

The constraint satisfaction (CS) fitness function is the average satisfaction of the Pareto-improving constraints and the incentive-compatibility constraints. $\Pi^i(j)$ is the cheating penalty when class i users switch to class j and $MinCP$ is the minimum cheating penalty in the current generation. The term $(SMC_i^1 - SMC_i^*)/SMC_i^1$ indicates the relative degree of welfare loss due to the transition where SMC_i^1 and SMC_i^* represent the full prices before and after the transition. The number of constraints is the sum of the number of Pareto-improving (PI) constraints (N') plus the number of incentive-compatibility (IC) constraints $N(N-1)$. Thus, the CS function ranges from 0 to 1 indicating the average amount of constraint satisfaction. The other two fitness functions discount the objective (changes in either net system value or revenue after the transition)[1] by the fraction of constraint satisfaction. Essentially, the fraction of constraint satisfaction provides a penalty to reduce the amount of the objective. Combining net system value and revenue in one fitness function is problematic because there is no obvious way to transform them to a comparable scale.

The remaining parts of the genetic algorithm are conventional (Goldberg 1989). We used mutation and single point crossover as the genetic operators. Both of these operators are applied to the base 10 representation of population members. We used roulette-wheel selection to select members of the current generation for genetic operations and actual implementation was coded using Maple 10.

5.2 Simulations

The simulations used random problems created by a sampling procedure. <Table 3> shows the range of values for service times, waiting costs, arrival rates, and coefficients of value functions used by the sampling procedure. Without loss of generality, the sampling procedure used value functions with derivatives of the form $V'(\lambda_i) = A_i - B_i\lambda_i$ (i = 1, 2) and exponential service time distributions.

<Table 3> Parameter Ranges for the Generated Problems

A_i	B_i	c_i	v_i
0 to 65	0 to 265	0 to 2	0 to 3

<Table 4> summarizes the generated problems by the number of classes. In order to put the algorithm in a more general setting, we extend the number of user classes from 2 to 4. We believe that it would be sufficient to provide up to 4 classes for differentiated Internet service for a couple of reasons. First the marginal benefit going from a non-priority system to a system with more priority classes will decline (Wilson 1989) and adding more prioritized service classes only leads to marginal revenue increase, which was observed in our simulation result. Second, we cannot expect prioritized services having more than 4 classes for all practical

1) Budget imbalance can be another consideration to the system manager as were argued by Dewan and Mendelson1990) and Mendelson (1985).

purposes. The Pareto Improving and Non Pareto Improving columns demonstrate that most post transition solutions are Pareto Improving although the ratio declines considerably (121 : 1 for 2 classes to 9 : 1 with 4 classes) with additional classes. This situation could be due to the increasing chance of having non-Pareto-improving classes with additional classes. To provide adequate test cases for the genetic algorithm, the Pareto Improving column includes Non Pareto Improving problems in which each worse-off class pays full price at least 1% more than before the prioritization. This threshold had a marginal impact on the results because there were only a few problems affected by the threshold. The Infeasible and Non Converging columns demonstrate the difficulty of solving this non-linear optimization problem. The Infeasible column means that the resulting solution had a negative value or its usage rate is greater than one. Negative values often mean that the resulting classes can be different after the transition because some classes are eliminated by the transition. In other words, we did not count the class dominance cases as was reported by Balachandran and Radhakrishnan (1994). The Non Converging column contains the number of problems in which the Maple 10 Solver failed to find a solution satisfying non-negativity and first order conditions for (1). Non-convergence was a significant problem in both the sampling procedure and the simulation.

〈Table 4〉 Summary of Generated Problems

Classes	Pareto Improving	Non Pareto Improving	Infeasible	Non Converging
2	24,298	200	1,129	4,138
3	5,548	200	245	3,093
4	1,976	200	62	2,010

Each simulation compared the genetic algorithm to random search using the generated non Pareto-improving problems as input. The parameters used in the genetic algorithm <Table 5> are consistent with values used in other studies of constrained optimization problems (Goldberg 1989; Michalewicz 1996). In the random search, the heuristic objective function was used along with Δ_i values randomly generated in a range determined by the gap between the two full prices (social marginal costs) of class-i users after and before the transition. The genetic algorithm used these randomly generated solutions as its initial population. The genetic algorithm was executed for each combination of a number of classes and a fitness function.

〈Table 5〉 Parameters for the Genetic Algorithm

Population Size	Crossover Rate	Mutation Rate	Number of Generations
10	0.6	0.1	40

The simulations demonstrated that the genetic algorithm provided only modest improvement over random search as shown in <Table 6> to <Table 8>. The hypothesis involves the percentage improvement of the best solution in 40 generations (i.e., terminal values) compared to the best solution in the initial population. In the fitness function columns, NSV represents the net system value while CS represents the constraint satisfaction (2). Each result column in Tables 6 to 8 contains less than 200 observations because of convergence problems when solving the Kuhn-Tucker conditions for the optimization problem (1). Although each hypothesis test strongly supports rejection of the null hypothesis, the sample

〈Table 6〉 Two Class Results of the Hypothesis H0 : % Improvement ≤ 0

	Fitness Function		
Z Test of Hypothesis for the Mean	*NSV* * *CS*	*Revenue* * *CS*	*CS*
Sample Standard Deviation	0.019424754	0.01607145	0.012815881
Sample Size	193	191	188
Sample Mean	0.005983098	0.00821336	0.003851304
Standard Error of the Mean	0.001398224	0.001162889	0.000934694
Z Test Statistic	4.279068819	7.062891041	4.120389419
Upper Critical Value (0.05 level)	1.644853	1.644853	1.644853
p-Value	9.39038E-06	8.20788E-13	1.89222E-05

〈Table 7〉 Three Class Results of the Hypothesis H0 : % Improvement ≤ 0

	Fitness Function		
Z Test of Hypothesis for the Mean	*NSV* * *CS*	*Revenue* * *CS*	*CS*
Sample Standard Deviation	0.039337018	0.0355581	0.022336994
Sample Size	149	184	160
Sample Mean	0.011099327	0.015391666	0.009295907
Standard Error of the Mean	0.003222614	0.002621378	0.001765894
Z Test Statistic	3.444199977	5.871593718	5.26413528
Upper Critical Value (0.05 level)	1.644853	1.644853	1.644853
p-Value	0.000286424	2.16519E-09	7.05677E-08

〈Table 8〉 Four Class Results of the Hypothesis H0 : % Improvement ≤ 0

	Fitness Function		
Z Test of Hypothesis for the Mean	*NSV* * *CS*	*Revenue* * *CS*	*CS*
Sample Standard Deviation	0.011607895	0.045513433	0.007981093
Sample Size	147	158	145
Sample Mean	0.005018496	0.009326635	0.00441466
Standard Error of the Mean	0.000957403	0.003620854	0.000662794
Z Test Statistic	5.241779525	2.575810592	6.660684623
Upper Critical Value (0.05 level)	1.644853	1.644853	1.644853
p-Value	7.96752E-08	0.0050003	1.37015E-11

means are rather small. In most cases, the mean percentage improvement is less than 1%. Z tests for one-sided tests with μ of 0.1 were accepted with p values of 1 for all cases.

The marginal contribution of the genetic algorithm may be explained by the small size of the transition externality compared to the total delay costs. Across all results, the average ratio of the difference of the total delay costs minus the transition externality to the total delay costs was

more than 0.90. If random search typically finds some improvement to reduce the transition externality, the genetic algorithm may not have enough flexibility for much additional improvement.

Another problem that may hinder the genetic algorithm is the diversity in the initial population. A more diverse initial population may provide more flexibility for the genetic algorithm. However, generating a more diverse initial population may significantly slow the optimization process due to weak convergence problem with Maple Solver. We are somewhat pessimistic on improving the genetic algorithm in this manner because the balance between convergence and diversity may be a harder problem than the original optimization problem.

6. Concluding Remarks

In this paper, we propose a solution for a network operator who, faced with unregulated growth in Internet traffic, may have to control his networks by offering prioritized services. Our proposal is based on a multi-class M/G/1 model that could serve as a good approximation of the network operator's already complex networks. Despite that previous studies in queueing congestion problems have demonstrated the benefits of prioritized queues, few of them have tried to analyze the change from the individual users' and system' perspectives. We showed welfare impact of prioritization on net system value and shifts in full prices and developed a model as well as a solution technique attaining Pareto-improving, optimal, and incentive-compatible states. Given the multiple constraints, we found that the exact solution is intractable and thus decided to resort to an approximate solution that can simultaneously satisfy the constraints while reaching an optimal state (out of many). Through simulation, we study the how constraints are satisfied under range of parameter values with approximate solutions. Although it could be deemed that the marginal improvement could be negligible as is illustrated in Example 2, we conclude that prioritization is more than a worthwhile effort, particularly for network operators whose revenue often runs in the amount of multi-billion dollars because even a fraction of billions still matters a lot.

There could be a number of future directions of the current study. Admittedly, the monopolist assumption offers just a baseline. As such, it would be interesting to extend our analysis into a more competitive market. What if other operators follow the suit? A more complex analysis would be needed to look into issues such as stability or sustainability of equilibrium under such conditions. For another, the assumption as to maximizing the net system value is too altruistic to be viewed as realistic and therefore the operator may seek for a solution favoring revenue increase in line with the recent interest in yield management. Of course, the lack of job class information may prevent a carrier from deploying the pricing models presented in this paper. However, with switching technologies enabling such prioritization at hands, the network carrier can try to improve its internal operation by trying and learning from the solution steps presented in this paper. The new challenges posed by VoIP, IPTV, Web 2.0 and P2P should give ample impetus to the network carrier who would like to improve its services.

Reference

[1] Balachandran, K. and S. Radhakrishnan, "Extensions to class dominance characteristics," *Management Science*, Vol.40, No.10 (1994), pp.353-360.

[2] Cocchi, R., D. Estrin, S. Shenker, and L. Zhang, "Pricing in computer networks : Motivation, formulation, and example," *IEEE/ACM Transactions on. Networking*, Vol.1, No.6(1993), pp.614-629.

[3] Dewan, S. and H. Mendelson, "User Delay Costs and Internal Pricing for a Service Facility." *Management Science*, Vol.36, No. 12(1990), pp.1502-17.

[4] Dovrolis, C. and P. Ramanathan, "A Case for Relative Differentiated Services and the Proportional Differentiation Model," *IEEE Network*, Vol.13, No.5(1999), pp.26-34.

[5] Edel, R. and P. Varaiya, "Providing Internet Access : What We Learn from INDEX," *IEEE Network*, Vol.13, No.5(1999), pp.18-25.

[6] Goldberg, D., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA, 1989.

[7] Hahn R. and Wallsten S., "The Economics of Net Neutrality," http://www.bepress.com /ev/vol3/iss6/art8/, 2006.

[8] Kim Y. and M. Mannino, "Optimal incentive-compatible pricing for M/G/1 Queues," *Operations Research Letters*, Vol.31, No.6(2003), pp.459-461.

[9] Kleinrock, L. *Queueing Systems, Vol. II : Computer Applications*, John Wiley and Sons, Inc., 1976.

[10] Knudsen, N. "Individual and Social Optimization in Multiserver Queue with a General Cost-Benefit Structure," *Econometric*, Vol. 40, No.3(1972), pp.515-528.

[11] Kwak, J.H., "On network neutrality issues in the US," http://www.kisdi.re.kr/image-data/pdf/10/1020061003.pdf.

[12] Laxton, Jr. W., "The End of Net Neutrality," www.law.duke.edu/journals/dltr/articles/2006dltr0015.html.

[13] Mendelson, H., "Pricing Computer Services : Queueing Effects," *Communications of the ACM*, Vol.28, No.3(1985), pp.312-321.

[14] Mendelson, H. and S. Whang, "Optimal Incentive-compatible Priority Pricing for the M/M/1 Queue," *Operations Research*, Vol.38, No.5(1990), pp.870-883.

[15] Michalewicz, Z., *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, Third Edition, Berlin, Germany, 1996.

[16] Mitrani I., *Probabilistic Modelling*, Cambridge University Press, 1998.

[17] Naor, P., "The Regulation of Queueing Size by Levying Tolls," *Econometrica*, Vol.37, No.1(1969), pp.15-24.

[18] Pigou, P., *The Economics of Welfare*, Macmillan, First Edition, London, 1920.

[19] Wilson, R., "Efficient and Competitive Rationing," *Econometrica*, Vol.57, No.1(1989), pp.1-40.