Generalized network engineering: Optimal pricing and routing for multi-service networks

Debasis Mitra Qiong Wang
Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974
{mitra,chiwang@lucent.com}

ABSTRACT

One of the functions of network engineering is to allocate resources optimally to forecasted demand. We generalize the mechanism by incorporating price-demand relationships into the problem formulation, and optimizing pricing and routing jointly to maximize total revenue. We consider a network, with fixed topology and link bandwidths, that offers multiple services, such as voice and data, each having characteristic price elasticity of demand, and quality of service and policy requirements on routing. Prices, which depend on service type and origin-destination, determine demands, that are routed, subject to their constraints, so as to maximize revenue. We study the basic properties of the optimal solution and prove that link shadow costs provide the basis for both optimal prices and optimal routing policies. We investigate the impact of input parameters, such as link capacities and price elasticities, on prices, demand growth, and routing policies. Asymptotic analyses, in which network bandwidth is scaled to grow, give results that are noteworthy for their qualitative insights. Several numerical examples illustrate the analyses.

Keywords: traffic engineering, routing, pricing, revenue maximization, capacity scaling, demand growth

1. INTRODUCTION

To fulfill its role as the universal communications infrastructure, the Internet is evolving to support a wide range of services at different levels of Quality of Service (QoS). Over the past several years, packet differentiation mechanisms like DiffServ have been proposed to allow IP traffic to receive specific per-hop forwarding behavior based on their QoS requirements.[4] It has also been recognized that besides these per-hop control mechanisms, network-wide planning and provisioning are also critical for satisfying end-to-end QoS guarantees. To this end, traffic engineering technologies, such as Multiprotocol Labeling Switching (MPLS), are under extensive study and development.[2, 3, 7]

A MPLS-type traffic engineering framework has two functions that apply to different time scales and complement each other.[3, 20] Given fixed network capacity, the network dimensioning function, which is based on demand forecasts, defines explicit routed paths and partitions capacities for different services. The dynamic route management function, which takes observed network states as inputs, employs real-time routing to balance loads on network links. While both are important to the delivery of QoS requirements, the focus of this paper is on the network dimensioning function.

Work on network dimensioning has concentrated on the tradeoff between satisfying QoS requirement and minimizing resources usage. The goal is to maximize carried demand, while minimizing the required network capacities. However, from a broader perspective, the higher level objective of a carrier is to maximize total revenue, which is earned by, first, generating appropriate amount of demands, and then, delivering them efficiently. Therefore, there is a need to generalize the model of network engineering, which subsumes dimensioning, to include both an optimal pricing policy to generate desired amounts of demands, and an efficient routing mechanism to map demands to network resources. This need is addressed in the paper.

Under this generalized network engineering model, the presence of multiple services has its respective influences on both routing and pricing. For each service, routing is constrained by its own QoS and policy requirements, while pricing is subject to distinct price-demand relationship, characterized by price elasticity of demand. Consequently, prices of different services as well as their respective representations in the total traffic are typically different. Both prices and routes need to be chosen to maximize total revenue, and there is a strong coupling between the two. For revenue to be earned, demand generated has to be matched with fixed network resources. Therefore, the optimal demand generation for various interacting multi-service traffic streams, as well as the allocation of network capacities to those streams, is strongly dependent on their unit prices. Based on the recognition of these dependencies and constraints, we develop, solve, and analyze a joint pricing and routing model for optimizing the network engineering process.

The time scale of our generalized network engineering model is intermediate between that of capacity planning and dynamic traffic engineering. The network, i.e., topology and link capacities, is fixed and determined by capacity planning. Conversely, price determination based on economic and business modeling makes the time scale longer than that of dynamic traffic engineering.[20]

The new model extends the scope of existing routing and traffic engineering studies, which usually take demands as exogeneously given, and exclude the influence of pricing.[3,13,17] Our results also contribute to the network economics research, where pricing schemes have typically been developed under the abstraction of a single-link network, in which routing is trivialized. Congestion pricing in networks is fundamentally separated from the subject of this paper by issues of demand modeling and time scales, although there are certain common underpinnings, such as shadow costs.[5,10,16,22] In particular, the modeling of congestion pricing typically includes variability in the demand process, which is sometimes done by postulating a Poisson process, as in.[11] In this paper, we do not model demand variability.

Importantly, the new model provides a framework for exploring issues related to the planning of future networks. For example, as the unit capacity cost is driven down by technology innovations, carriers respond by expanding their networks.[14] To make use of additional capacities, more demand needs to be generated by reducing prices, and routing needs to be redesigned to accommodate new demand profiles and increased capacities. Understanding the evolution of the optimal pricing and routing in an expanding network helps carriers to adopt appropriate controlling mechanisms for the long run. We present results which contribute to this understanding.

Our analysis is in two main parts. First, we formulate and solve the generalized network engineering model as a mathematical optimization problem. We establish connections between routing and pricing through link shadow costs, which are the Lagrange multipliers corresponding to the link capacity constraints in the optimization model. We prove the important result that in the optimal solution, these quantities should be used both as proxies of marginal cost for setting prices and directives for routing demands between node-pairs. Second, we analyze the optimal solution in an asymptotic framework. We derive upper and lower bounds for shadow costs, and use these bounds to explore the relationship between the price elasticity of demand, and routing policies. We infer properties of optimal solutions for large networks, in which link bandwidths are scaled to grow. We study the impact of the scaling on optimal prices and demands for various services, and convergence of the optimal routing policy. Our analytical results are backed by several numerical case studies.

The organization of the paper is as follows. We develop and solve the optimization model in Section 2, derive analytical results from asymptotic analyses in Section 3, and present numerical examples in Section 4. In Section 5, we conclude the paper by outlining major results and future directions.

2. PROBLEM FORMULATION AND BASIC SOLUTION PROPERTIES

In this section, we develop a generalized network engineering model for maximizing a carrier's revenue in a multi-service network. In 2.1, we discuss two pivotal concepts of our model: price-elasticity of demand and admissible route sets. In 2.2, we develop mathematical formulations of the model that jointly optimize price and routing decisions. In 2.3, we follow Lagrange's method for concave programming, and show that the shadow costs are effective devices for implementing both optimal pricing and routing.

2.1. Price elasticity and admissible route sets

Modeling a multi-service network starts with modeling distinct features of each service, including both price-demand relationship and QoS requirements. First, for every particular service, we define $D = D(P)$ as the demand function, where D denotes demand and P denotes price. An essential parameter is the price-elasticity of demand, defined as *:

$$\epsilon = -(\frac{P}{D})\frac{dD}{dP} \tag{1}$$

A larger value of price elasticity for a service indicates that a small percentage decrease in price generates more demand of that service. Furthermore, revenue, denoted by R, is given by the product of price and demand, i.e. $R = PD$. It can be verified that depending on whether $\epsilon > 1$, $\epsilon = 1$, or $\epsilon < 1$, reduction in price leads to increase, no change, or reduction in revenue. The larger the elasticity, the more revenue is generated by reducing price.

While the analytic framework of this paper is sufficiently general to accommodate demand models of many forms, previous studies indicate that constant price-elasticity model is appropriate for bandwidth. Hence we will assume that (1) holds for all P and D with ϵ held fixed. Hence, for $P > 0$,

$$D = \frac{A}{P^{\epsilon}}. \tag{2}$$

The constant A is equal to the value of D when $P = 1$, and it may be interpreted as demand potential. It has also been estimated[14,15] that price-elasticity of voice traffic is approximately 1.05, much lower than that of data service, which is in the range of 1.3-1.7. Throughout this paper, we assume that $\epsilon > 1$.

QoS requirements are another defining characteristic of a service. We rely upon two concepts for the planning of the delivery of QoS, namely, effective bandwidth of individual flows, or connections, which depends on the connection's service type, and admissible route sets. For service s and (source, destination) pair (σ_1, σ_2), the total bandwidth for $(s, (\sigma_1,\sigma_2))$ is given by the product of the effective bandwidth of individual connections of the service and the rate of connections. (We abbreviate (σ_1,σ_2) to σ, and refer to (s,σ) as a *stream*). The network engineering problem deals with the total stream bandwidth, and consequently the notion of effective bandwidth, while obviously important, does not feature directly in the analysis.

Admissible route sets, on the other hand, are ubiquitous in this paper. It is a device that allows us to take into account constraints on routing imposed by QoS and, more generally, policy considerations. To handle end-to-end constraints of a service, we let $\mathcal{R}(s,\sigma)$ denote the set of admissible routes for the stream (s,σ). For example, real-time services, such as voice and video, may require routes with lengths not exceeding specified thresholds, on account of propagation delay, and, typically, restrictions will also apply on the number of hops in the routes, since each hop is associated with an additional switching node and consequent incremental delay. In contrast, for delay insensitive service classes, the distance and hop constraint would be less stringent. Importantly, the admissibility of a route will also depend on policy, which might reflect diverse considerations, such as security, the capability of switching nodes in the routes to handle certain services, link capacity, etc. Generating the admissible route sets, $\mathcal{R}(s,\sigma)$, is a substantial task in itself. However, in this paper we assume that these sets are given.

In Section 4, where we present results from our numerical investigations, the only restriction that the admissible routes are required to satisfy is on the maximum number of hops, a service-specific parameter. This criterion is just for illustrative purposes.

We make the simplifying assumption that for σ_1 and σ_2 which are adjacent, i.e., a direct link connects σ_1 to σ_2, the admissible route set $\mathcal{R}(s,\sigma)$ always includes the direct link for each service. That is, the direct link is always an admissible route for its end points.

2.2. Model for revenue maximization

Let s $(s = 1, 2 \ldots, S)$ denote service class. Let (σ_1, σ_2) denote (source, destination) pair, and (s, σ) a stream. $P_{s\sigma}$ is the price for unit bandwidth for stream (s, σ). $D_{s\sigma}$ is the bandwidth demand for stream (s, σ). The (fixed) capacity, in units of bandwidth, of link l, is denoted by C_l $(l = 1, 2, \ldots, L)$. For $r \in \mathcal{R}(s, \sigma)$, we call (s, r) a service route, and let X_{sr} denote the carried bandwidth on this service route. Except in Section 3, where

*Here, we consider point elasticity as opposed to arc elasticity, see[12] for the difference between the two.

specialized results are obtained, we allow distinct constant price elasticity models for each stream (s, σ). That is, for each (s, σ),

$$D_{s\sigma} = \frac{A_{s\sigma}}{P_{s\sigma}^{\epsilon_{s\sigma}}} \quad \text{or} \quad P_{s\sigma} = (\frac{A_{s\sigma}}{D_{s\sigma}})^{1/\epsilon_{s\sigma}} \tag{3}$$

where $\epsilon_{s\sigma}$ and $A_{s\sigma}$ are respectively the elasticity and demand potential for the stream. As previously noted, we assume that $\epsilon_{s\sigma} > 1$, for all (s, σ).

The revenue generated on a stream depends on the product of the carried bandwidth on the stream and the unit price for bandwidth for the stream. The network revenue, W, is obtained by summing over all streams. The revenue maximization problem is:

$$\max_{\{P_{s\sigma}\},\{X_{sr}\}} W = \sum_{s,\sigma} P_{s\sigma} \sum_{r \in \mathcal{R}(s,\sigma)} X_{sr} \tag{4}$$

subject to the following constraints:

$$\sum_{r \in \mathcal{R}(s,\sigma)} X_{sr} \leq D_{s\sigma} \qquad \forall (s, \sigma) \tag{5}$$

$$\sum_{(s,\sigma)} \sum_{r \in \mathcal{R}(s,\sigma): l \in r} X_{sr} \leq C_l \qquad \forall l \tag{6}$$

$$P_{s\sigma} \geq 0 \qquad \forall (s, \sigma) \tag{7}$$

$$X_{sr} \geq 0 \qquad \forall (s, \sigma), \forall r \in \mathcal{R}(s, \sigma) \tag{8}$$

Equation (5) states that for each stream, the total carried bandwidth does not exceed the demand, and (6) reflects link capacity constraints. In the above formulation, the price and demand variables are both represented, even though these variables are related through the demand model in (3). We will find it convenient to eliminate the price variables and retain only the demand variables. The optimum prices may, of course, be recovered from the demands in the solution. Transforming the objective function gives:

$$\max_{\{D_{s\sigma}\},\{X_{sr}\}} W = \sum_{(s,\sigma)} (\frac{A_{s\sigma}}{D_{s\sigma}})^{1/\epsilon_{s\sigma}} \sum_{r \in \mathcal{R}(s,\sigma)} X_{sr} \tag{9}$$

subject to the constraints in (5) and (6), and the non-negativity of $D_{s\sigma}$ and X_{sr}.

We next note that the optimal solution will satisfy the constraint in (5) with equality. The reason is that for any stream (s, σ) for which there is slack in (5), we may hold fixed all X_{sr} for the stream, so that the total carried bandwidth on the stream is fixed, while decreasing $D_{s\sigma}$ to the point where the slack disappears, by increasing price $P_{s\sigma}$ appropriately. Hence, revenue from stream (s, σ) is increased, without there being any impact on other streams.

By related reasoning, it can also be shown that the optimal solution will satisfy (6) with equality. This is a consequence of the assumption (see Section 2.1) that links are admissible routes for its end-points for each service. We argue that the demand and the carried bandwidth on the direct link may be increased equally by an amount that makes the unused capacity on the link vanish. The increase in demand is achieved by dropping the price, which, by virtue of the assumption that elasticity exceeds unity, leads to increased revenue.

With this observation, the joint pricing and routing problem becomes:

$$\max_{\{D_{s\sigma}\},\{X_{sr}\}} W = \sum_{s,\sigma} A_{s\sigma}^{1/\epsilon_{s\sigma}} D_{s\sigma}^{(\epsilon_{s\sigma}-1)/\epsilon_{s\sigma}} \tag{10}$$

subject to

$$\sum_{r \in \mathcal{R}(s,\sigma)} X_{sr} = D_{s\sigma} \qquad \forall (s, \sigma) \tag{11}$$

$$\sum_{s,\sigma} \sum_{r\in\mathcal{R}(s,\sigma):l\in r} X_{sr} = C_l \quad \forall l \tag{12}$$

$$D_{s\sigma} \geq 0 \qquad \forall (s,\sigma) \tag{13}$$

$$X_{sr} \geq 0 \qquad \forall (s,\sigma) \text{ and } \forall r \in \mathcal{R}(s,\sigma) \tag{14}$$

Note that as a consequence of $\epsilon_{s\sigma} > 1$, $\forall (s,\sigma)$, the objective function is a concave, monotonically increasing, differentiable function of $D_{s\sigma}$, and that the constraints are linear in the decision variables. The above is a special case of the class of "Concave Programming" problems, which have attractive properties.[6] Also, effective algorithms exist for concave programming.[21]

2.3. Shadow costs and minimum cost routing

We follow Lagrange's Method[6, 9] in concave programming. Define the Lagrangian:

$$\mathcal{L}(\boldsymbol{D},\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\lambda}) = \sum_{s,\sigma} A_{s\sigma}^{1/\epsilon_{s\sigma}} D_{s\sigma}^{(\epsilon_{s\sigma}-1)/\epsilon_{s\sigma}} + \sum_{s,\sigma} \mu_{s\sigma}\Big(\sum_{r\in\mathcal{R}(s,\sigma)} X_{sr} - D_{s\sigma}\Big) + \sum_{l} \lambda_l\Big(C_l - \sum_{s,\sigma} \sum_{r\in\mathcal{R}(s,\sigma):l\in r} X_{sr}\Big) \tag{15}$$

Here $\{\mu_{s\sigma}\}$, $\{\lambda_l\}$ are the Lagrange multipliers associated respectively with the constraints implied by demand satisfaction in (11) and the link capacities in (12).

For the Kuhn-Tucker Theorem to hold, we require the constraint qualification, sometimes called the Slater condition, to hold. This requires that the matrix $\boldsymbol{G}$ is of maximum rank, where $\boldsymbol{G}$ is defined to be such that its application to the vector of decision variables ($\boldsymbol{D}$,$\boldsymbol{X}$) gives the left hand side of the constraints in (5) and (6). It can be shown that provided the admissible route sets satisfy simple non-degeneracy conditions, the constraint qualification is satisfied.

A main result from Concave Programming states that if ($\boldsymbol{D}$,$\boldsymbol{X}$) maximizes revenue W, then there exists ($\boldsymbol{\mu}$,$\boldsymbol{\lambda}$), such that

1. ($\boldsymbol{D}$,$\boldsymbol{X}$) maximizes $\mathcal{L}$ in (15) without any constraints, i.e. for all streams (s,σ)

$$\frac{\partial \mathcal{L}}{\partial D_{s\sigma}} \leq 0,\ D_{s\sigma} \geq 0 \quad \text{with complementary slackness,} \tag{16}$$

$$\frac{\partial \mathcal{L}}{\partial X_{sr}} \leq 0,\ X_{sr} \geq 0 \quad \text{with complementary slackness,} \tag{17}$$

2. $\mu_{s\sigma} > 0 \quad \forall (s,\sigma) \qquad \text{and} \qquad \lambda_l > 0 \quad \forall l$

The Lagrange multipliers $\{\mu_{s\sigma}\}$ and $\{\lambda_l\}$ are shadow costs with properties that are discussed next. First, from (16) and (17) we have the following important relationship between the shadow cost $\mu_{s\sigma}$ and the optimal price $P_{s\sigma}$ for the stream (s,σ):

$$P_{s\sigma} = \Big(\frac{A_{s\sigma}}{D_{s\sigma}}\Big)^{1/\epsilon_{s\sigma}} = \frac{\epsilon_{s\sigma}}{\epsilon_{s\sigma}-1}\mu_{s\sigma} \quad \forall (s,\sigma) \tag{18}$$

Next, for each (s,σ) and for each admissible route r for which the (source, destination) is σ, we have:

$$\text{either} \quad X_{sr} > 0 \quad \text{and} \quad \sum_{l\in r} \lambda_l = \mu_{s\sigma} \tag{19}$$

$$\text{or} \quad X_{sr} = 0 \quad \text{and} \quad \sum_{l\in r} \lambda_l \geq \mu_{s\sigma} \tag{20}$$

The above condition is of considerable interest. To examine its consequences, begin by interpreting $\mu_{s\sigma}$ as the *minimum route cost* for (s,σ). Next, we interpret λ_l as the *link cost*. Finally, for any route r, we define the *route cost* to be the sum of link costs, i.e., $\sum_{l\in r}\lambda_l$. Thus, the condition in (19) states that if in the optimal solution an admissible route r for the stream (s,σ) is "active", in the sense that it carries traffic, then necessarily the route cost equals the minimum route cost for the stream. In other words, traffic of each stream (s,σ) can only be carried on routes that belong to the *minimum cost admissible route set* $M(s,\sigma)$, defined as:

$$M(s,\sigma) \equiv \{r | r \in \mathcal{R}(s,\sigma) \cap \sum_{l \in r} \lambda_l = \mu_{s\sigma}\}, \tag{21}$$

which is typically a considerably thinned subset of the set of admissible routes. Thus, from the properties of the Concave Programming solution we have obtained link costs and a "minimum cost routing" policy, which is optimal.

We end this section by emphasizing two points: first, an intimate relation exists between optimal stream prices, $\{P_{s\sigma}\}$, and the minimum route costs for streams $\{\mu_{s\sigma}\}$, and, second, link costs do not depend on the service type, which simplifies the implementation of the optimal routing policy.

3. AYMPOTOTIC PROPERTIES OF THE OPTIMAL SOLUTION

In this section, we investigate asymptotic properties of optimal solutions, specifically, the influences of price elasticity and uniform capacity expansion.

By "uniform capacity expansion", we mean the scaling in which the capacities of all links grow in proportion to a common multiplier. Such expansions occur, for instance, when a carrier replaces existing systems with equipment of a later generation. For example, capacities of all network links will grow uniformly by a factor of 4 when 2.5 Gbits/sec optical transmission systems are replaced by 10 Gbits/sec systems. Now, recall from Section 2.1 that new data services are more elastic with respect to price changes than the traditional voice service. The combined effects of different price elasticities and uniform capacity expansion are important in multi-service networks. We examine their impacts on the optimal prices and demands, and on the optimal routing.

3.1. Bounds on shadow costs

We establish upper and lower bounds for link shadow costs, which are used to prove several key properties later in the paper. These results are limited to the case where $\epsilon_{s\sigma} = \epsilon_s$, for all σ, i.e., the price elasticity depends on the service type, but *not* on the (source, destination). The following notation is used:

$$\epsilon_{max} = \max_s \epsilon_s, \ \epsilon_{min} = \min_s \epsilon_s, \qquad C_{min} = \min_l C_l, \ C_{max} = \max_l C_l,$$
$$A_{s,min} = \min_\sigma A_{s\sigma}, \ A_{s,max} = \max_\sigma A_{s\sigma}, \quad \tilde{A}_{s\sigma} = (\tfrac{\epsilon_s - 1}{\epsilon_s})^{\epsilon_s} A_{s\sigma}, \ \tilde{A}_{s,min} = \min_\sigma \tilde{A}_{s\sigma} \ \tilde{A}_{min} = \min_s \tilde{A}_{s,min} \tag{22}$$

Lemma 1 (see Appendix for proof)

Let λ_l be the optimal shadow cost of link l. Then for any $s \in S$ and $l \in L$:

$$(\frac{\tilde{A}_{s,min}}{C_{max}})^{1/\epsilon_s} \leq \lambda_l \leq \max[(\frac{\sum_{s,\sigma} \tilde{A}_{s\sigma}}{C_{min}})^{1/\epsilon_{max}}, (\frac{\sum_{s,\sigma} \tilde{A}_{s\sigma}}{C_{min}})^{1/\epsilon_{min}}] \tag{23}$$

For large values of price elasticity for *all* services, the difference between the upper and lower bounds of shadow costs in (23) will diminish. For example, if $C_l \leq \sum_{s,\sigma} \tilde{A}_{s\sigma}$ for all l:

$$(\frac{\tilde{A}_{min}}{C_{max}})^{1/\epsilon_{min}} \leq \lambda_l \leq (\frac{\sum_{s,\sigma} \tilde{A}_{s\sigma}}{C_{min}})^{1/\epsilon_{min}} \tag{24}$$

The ratio of the lower bound to the upper bound, $(\frac{C_{min}\tilde{A}_{min}}{C_{max}\sum_{s,\sigma} \tilde{A}_{s\sigma}})^{1/\epsilon_{min}}$, is monotonically increasing in ϵ_{min}, and approaches 1 at its limit. Furthermore, when network capacity becomes sufficiently large, it only requires price elasticity for *some* services to be large to reduce the difference between two bounds. For example, if

$\sum_{s,\sigma} \tilde{A}_{s\sigma} < C_l$ for all l, in which case the upper bound of λ_l is $(\frac{\sum_{s,\sigma} \tilde{A}_{s\sigma}}{C_{min}})^{1/\epsilon_{max}}$, the ratio of the lower bound to the upper bound is $(\frac{C_{min}\tilde{A}_{min}}{C_{max}\sum_{s,\sigma} \tilde{A}_{s\sigma}})^{1/\epsilon_{max}}$, which increases with ϵ_{max}. As the gap between the lower and the upper bounds shrinks, the optimal shadow costs become increasingly uniform across all links.

Why does high price elasticity have the tendency to equalize the optimal link shadow costs? To answer this question, start with a scenario where all shadow costs are set to the uniform upper bound, $(\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min})^{1/\epsilon_{max}}$, which is usually above their optimal values. From (3) and (18), overestimating shadow costs keeps demands below their optimal values, and underuses network capacities. To reach the optimum, shadow costs of some links have to be reduced from the upper bound to increase demands to the point where all link capacities are fully utilized. The extent to which shadow costs are allowed to decrease depends on the price elasticity. With high elasticity, small decreases in prices lead to larger increases in demands. Therefore, shadow costs reduction is limited, as otherwise more demands would be induced than can be handled by the network. Consequently, the lower bound of shadow costs stays closer to the upper bound, and the optimal shadow costs take more uniform values.

The aforementioned phenomenon and the discussion should help the reader reconcile some of the findings in the case studies reported in Section 4. In particular, uniformity of link shadow costs implies that minimum cost routes are ones with minimum number of hops.

3.2. Asymptotic properties of optimal prices and demands

The bounds on the optimal link shadow costs in Lemma 1, are used to obtain the optimal shadow costs under uniform capacity expansion in the following theorem.

Theorem 3-1 (see Appendix for proof)

Let $\lambda_l(1)$ be the optimal shadow costs for some base capacity C_{l0} $(l = 1, 2, \ldots, L)$, and $\lambda_l(m)$ be the optimal shadow costs when $C_l = mC_{l0}$, where m is a common scaling factor. Let $C_{max0} = \max_l C_{l0}$ and $C_{min0} = \min_l C_{l0}$. Suppose $m > 1$ and $C_{min0} \geq \sum_{s,\sigma} \tilde{A}_{s\sigma}$, where $\tilde{A}_{s\sigma}$ is defined in (22). Then:

$$\frac{m^{-1/\epsilon_{max}}}{G} \leq \frac{\lambda_l(m)}{\lambda_l(1)} \leq G m^{-1/\epsilon_{max}} \quad \textit{where } G = (\frac{C_{max0}\sum_{s,\sigma} \tilde{A}_{s\sigma}}{C_{min0}\tilde{A}_{min}})^{1/\epsilon_{max}} \textit{ is independent of } m \tag{25}$$

Combining Theorem 3-1 with the results in Section 2.3, we arrive at the following qualitative behavior of optimal pricing and demand under uniform capacity expansion.

Corollary :

1. Since the optimal prices for all services and node pairs are linear combinations of link shadow costs (see (18), (19), and (20)), $P_{s\sigma}(m)/P_{s\sigma}(1) = O(m^{-1/\epsilon_{max}})$, as $m \to \infty$, i.e., uniformly expanding network capacity by a factor of m reduces optimal prices by a factor of the order of $m^{1/\epsilon_{max}}$. Because $0 < 1/\epsilon_{max} < 1$, the rate of the price decrease is lower than the rate of capacity increase. The greater the highest price elasticity, the slower is the decrease of prices.

2. From the price-demand relationship, (see Section 2.1), it follows that,

$$\frac{D_{s\sigma}(m)}{D_{s\sigma}(1)} = \frac{A_{s\sigma}P_{s\sigma}^{-\epsilon_s}(m)}{A_{s\sigma}P_{s\sigma}^{-\epsilon_s}(1)} = O(m^{\epsilon_s/\epsilon_{max}}), \tag{26}$$

   which implies that under uniform capacity expansion, only demands with the highest price elasticity grow linearly with the capacity. Demands for other services grow at sub-linear rates determined by their price elasticity.

The above indicates that uniform capacity expansion will ultimately make the demand for the service with the highest price-elasticity the dominant component of total traffic. Interestingly, we can also derive a result on the dominance in traffic composition within a service class. As shown below in Theorem 3-2, if a service has sufficiently high price elasticity, then, in the optimal solution, demands between neighboring nodes will dominate total demands of that service.

Theorem 3-2 (see Appendix for proof)

Let D_s^d be the total demand of service s that is carried on direct links, i.e.$D_s^d = \sum_{\sigma : M(s,\sigma) \subset L} D_{s\sigma}$. Then

$$\lim_{\epsilon_s \to \infty} D_s^d / \sum_{\sigma} D_{s\sigma} = 1 \qquad (27)$$

3.3. Asymptotic properties of optimal routing

We are interested in knowing whether the uniform capacity expansion simplifies the optimal routing. For example, does the minimum cost admissible route sets contain only routes with the mininum number of hops? We would also like to know whether, for the sake of easier implementation,[8, 19] each stream can be carried on a single path, rather than be split across multiple routes? Our findings are summarized in the following three statements:

1. Uniform capacity expansion does not necessarily result in minimum-hop or non-split routing.
2. In a high capacity network, high price elasticity of at least one service leads to minimum-hop routing for all services.
3. High price elasticity does not necessarily result in non-split routing, but the total amount of split traffic is reduced.

The proof of the first statement is based on an example of a three-node network shown in Figure 1. Suppose that only one service, with price elasticity ϵ, is offered. Let $C_{a-b} = C_{b-c} = C_{a-c} = \bar{c}$, $A_{a,b} = A_{b,c} = \bar{A}$, and $A_{a,c} = 4^{\epsilon}\bar{A}$. Assume all routes of one or two hops are admissible. If routing is restricted only to routes with minimum number of hops and a single path for each node pair, then all demands, $D_{(a,b)}$, $D_{(b,c)}$, and $D_{(a,c)}$, will be carried on direct links $a-b$, $b-c$, and $a-c$, respectively. Link shadow costs can then be calculated from (11)-(14), (18)-(20) as follows:

$$\lambda_{a-b} = \lambda_{b-c} = \frac{\epsilon - 1}{\epsilon}(\frac{\bar{A}}{\bar{c}})^{1/\epsilon}, \quad \lambda_{a-c} = \frac{4(\epsilon - 1)}{\epsilon}(\frac{\bar{A}}{\bar{c}})^{1/\epsilon} \qquad (28)$$

Since $\lambda_{a-b} + \lambda_{b-c} < \lambda_{a-c}$, the route that carries demand $D_{(a,c)}$, i.e., link $a-c$, is not of the minimum cost, so the solution is not optimal. More specifically, total revenue in this case is $6(\bar{A}/\bar{c})^{1/\epsilon}\bar{c}$. The optimal revenue, which is achievable only when demand $D_{(a,c)}$ is carried by both routes $a-c$ and $a-b-c$ (so both minimum-hop and single path routing restrictions are violated), is $4[(1+2^{\epsilon})/2]^{1/\epsilon}(\bar{A}/\bar{c})^{1/\epsilon}\bar{c}$. The ratio between the two revenues is $3/2[(1+2^{\epsilon})/2]^{1/\epsilon}$, which equals 97% when $\epsilon = 1.5$. The fraction remains constant when $C_l = m\bar{c}$ for any $m > 1$.

The example demonstrates that the minimum-hop and single-path routing can lead to lower revenue than the optimal, and this is so regardless of the extent to which network capacities are ultimately increased. The second statement is based on the following theorem.

Theorem 3-3 (see Appendix for proof)

Let r_1, r_2 be any two routes connecting a given node pair, with $H(r_1)$, $H(r_2)$ being their respective hop counts. Suppose $C_{min} \geq \sum_{s,\sigma} \tilde{A}_{s\sigma}$, and let λ_l be the optimal shadow cost for link l. Then:

$$\exists \epsilon_m > 0, \text{ such that if } \epsilon_{max} \geq \epsilon_m \text{ then } H(r_1) > H(r_2) \Rightarrow \sum_{l \in r_1} \lambda_l > \sum_{l \in r_2} \lambda_l \qquad (29)$$

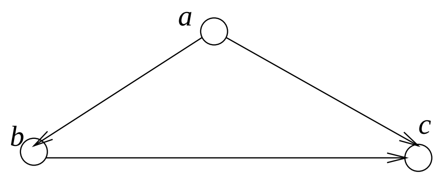

Figure 1. Example of Optimal Routing Under Uniform Capacity Expansion

Theorem 3-3 establishes the important result that any non-minimum-hop route cannot be a minimum cost route in a network with sufficient high capacity where, additionally, at least one service has sufficiently high price elasticity. In such a network, minimum-hop routes will be the optimal routes for demands between every node pair and every service.

The third statement is again based an example. Consider a four-node network(see Figure 2) with five pairs of demands, and assume $A_{a,b} = A_{b,c} = A_{a,d} = A_{d,c} = A_{a,c}$. The admissible route sets is $a-b-c$, $a-d-c$ for node pair (a,c), and a direct link for all others. When the restriction of non-split routing applies, demand from

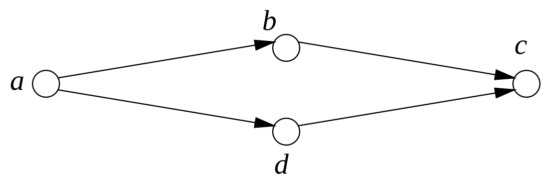

Figure 2. Example of Traffic-Splitting under High Price Elasticity

a to c is carried entirely either on route $a-b-c$ or $a-d-c$. In this case, regardless of the price elasticity, the shadow cost of the route that carries demand $D_{(a,c)}$ will always be higher than shadow costs of the alternate route that doesn't carry it. Therefore, non-split routing cannot be optimal.

Nevertheless, as shown in Theorem 3-2, in a high capacity network, the total traffic of a service with sufficiently high price elasticity will be dominated by demands between neighboring nodes, which, according to Theorem 3-3, is optimally routed directly and without splitting. Hence, even if high price elasticity will not make non-split routing optimal, it has the effect of reducing the total amount of split traffic.

4. CASE STUDIES

4.1. Sample network and base case

We consider a sample network shown in Figure 3, which has eight nodes and 20 directed links. The node and link indices are indicated in the figure. The network offers two different services, say voice and data, indexed by $s = 1, 2$, respectively. Assume that the admissible route sets are as follows: the voice service can only be carried on minimum-hop routes, while the data service may be carried on routes with up to five hops.

In the constant price elasticity formulation of the demand function for each service(see Section 2.1), we make the assumption here that the demand potential $A_{s\sigma}$ may depend on both s and σ, while price elasticity depends only on service type, i.e. $\epsilon_{s\sigma} = \epsilon_s$.

We construct a base case by letting $C_l = 400$ for all l, and $A_{1\sigma} = 2000$, and $A_{2\sigma} = 200$ for all σ, and $\epsilon_1 = 1.05$; $\epsilon_2 = 1.5$. We vary these parameters in the following studies.

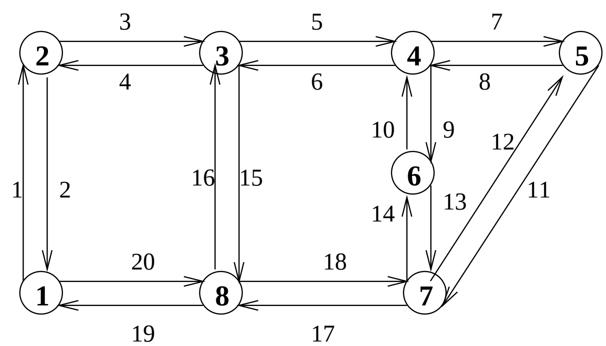

Figure 3. Sample Network

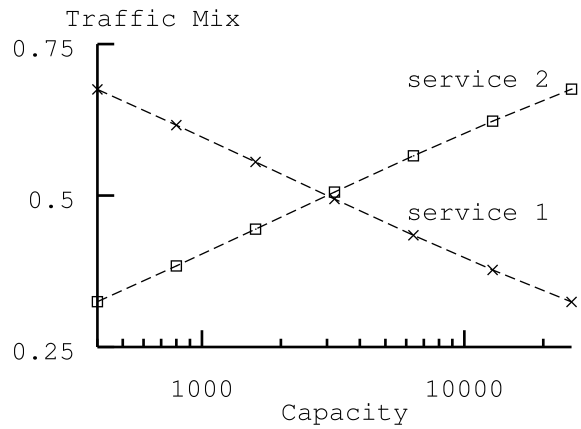

Figure 4. Change of Traffic Mix

4.2. Results

4.2.1. Effects of uniform capacity expansion

We first consider the case of uniform capacity expansion, and examine the resulting changes in the traffic composition. We let C_l start at 400, and in each step we double it till the capacity of 25600 is reached. In Figure 4, the horizontal axis is capacity in log-scale.

For each service, we sum demands for all (source, destination) pairs, and divide the sum by the total amount of traffic in the network, i.e., $\dfrac{\sum_\sigma D_{i\sigma}}{\sum_s \sum_\sigma D_{s\sigma}}$, $i = 1, 2$. The two lines in the figure show the respective ratios for the two services. At the lowest capacity level, about 70% of total traffic is of service 1. The ratio declines when capacity increases, and is down to a value just above 30% when the capacity reaches 25600. In contrast, the percentage of service 2 traffic rises from about 30% to 70%. This result is expected based on the analysis presented earlier (see (26)). Since price elasticity of service 2 is higher than that of service 1, the amount of service 2 traffic grows linearly with capacity, while that of service 1 grows sub-linearly, i.e., as the capacity raised to the power of ϵ_1/ϵ_2, which in this example is $1.05/1.5 = 0.7$. Therefore, as we keep doubling link capacities, data streams (service 2) gradually dominate voice streams (service 1) in the total traffic bandwidth. The change of traffic mix influences both the revenue and link shadow prices. Recall from our earlier discussion in Section 2.1 on demand functions that adding a unit of capacity will generate more incremental revenue if it is used to carry flows with higher price elasticity. Following this logic, one may conjecture that in this example, as link capacities increases, and service 2 traffic becomes dominant, the revenue growth should accelerate. Table 1 gives the percentage increase of revenue obtained from doubling capacity at different capacity levels. The conjecture is true here, as indicated by the higher growth rate at a higher capacity level. In a single-service

Table 1. Revenue Growth Rate

capacity	revenue growth rate
400	104.67%
800	105.03%
1600	105.46%
3200	105.98%
6400	106.59%
12800	107.31%

network, when the network capacity is increased by a factor of m, the shadow cost decreases exactly by a factor of $m^{1/\epsilon}$. Though this rule does not apply to a multi-service network, the difference is expected to diminish as network capacity increases. This is because increasing network capacity causes streams of the service with the highest price elasticity to be a dominant component of total traffic. Consequently, the evolution of shadow costs in those networks should resemble that of single-service networks, which offer only that service. To demonstrate this point, we compare shadow costs, $\lambda_l(2C)$ at capacity level $2C$, to $2^{-1/\epsilon_2}\lambda_l(C)$. Let e_l be the percentage difference, i.e., $e_l = (|\lambda_l(2C) - 2^{-1/\epsilon_2}\lambda_l(C)|)/\lambda_l(C)$, and Table 2 shows the maximum and the average value of e_l over all l. As expected, both the maximum and average percentage differences decrease with the capacity.

Table 2. Evolution of Link Shadow Costs

capacity	max(e_l)	mean(e_l)
400	14.5%	12.0%
800	13.1%	10.9%
1600	11.3%	9.0%
3200	10.4%	8.0%
6400	9.2%	7.6%
12800	8.2%	6.1%

4.2.2. Effects of price elasticity

The next example focuses on the effect of price elasticity. The parameters for this example are the same as in the base case, except that we let $A_{1,(7,5)} = 10000$, a large number. Because of the minimum-hop constraint, the flow of service 1 from node 7 to node 5 has to be carried on the direct link (link 12). Therefore, assigning a large value to $A_{1,(7,5)}$ makes link 12 a bottleneck with a high link shadow cost. We investigate whether other streams will take a longer route as a consequence.

Specifically, we consider a stream from node 1 to 5 (stream A) and a stream from node 8 to 5 (stream B). Both streams are of service 2, and their admissible routes are allowed to take non-minimum-hop ones. We consider five different cases in which ϵ_2 takes values of 1.1, 1.2, 1.3, 1.4, and 1.5, where ϵ_1 is fixed at 1.05. As shown in Figure 5, in all these cases, stream $A(B)$ is carried on a minimum-hop route, $R\_A(R\_B)$, or a non-minimum-hop route, $r\_A(r\_B)$. Each route is shown as a dashed line in the figure, and the percentages of stream A or B allocated to a route are also given in Table 3. When the elasticity ϵ_2 is small, all streams A and B are carried on non-minimum-hop routes r_A and r_B, respectively. As we discussed in Section 3.1, when elasticity increases, shadow costs become increasingly uniform across all links, so minimum-cost routes will become those with minimum number of hops. As a result, the two streams migrate to the minimum-hop routes, R_A and R_B, and eventually are completely carried by them. The example confirms our conclusion in Section 3.3 that high price elasticity induces minimum-hop routing.

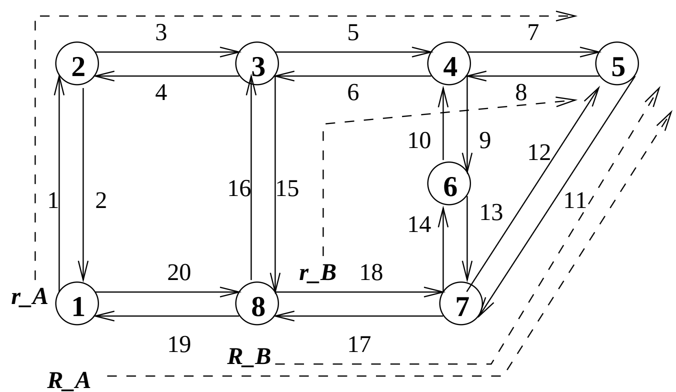

Figure 5. Min-Hop Routing Implied by High Price Elasticity

Table 3. Optimal Routing Under Different Price Elasticity

ϵ_2	route r_A	route R_A	route r_B	route R_B
1.1	100%	0%	100%	0%
1.2	100%	0%	28%	72%
1.3	62.7%	37.3%	0%	100%
1.4	9.5%	90.5%	0%	100%
1.5	0%	100%	0%	100%

4.2.3. Effects on route splitting

In Section 3.3 we proved that in some networks, traffic-splitting is inevitable for revenue maximization. We now illustrate through examples how uniform capacity expansion and price elasticity affect traffic splitting. The framework here is the same as in the base case, except that $A_{1\sigma} = 50$. The purpose of resetting $A_{1\sigma}$ is to create a starting case in which a significant portion of total traffic is carried on multiple routes. The ratio of split traffic, calculated by dividing the total amount of split traffic by the total traffic, is 21.44%. We then vary price elasticity, ϵ_2, from 1.1 to 1.9 in increments of 0.1. We also double network capacity uniformly from the base value of 400 to 25600. Figure 6 shows the split ratio under different combinations of price elasticity and link capacities. As shown in the figure, uniform capacity expansion has little impact on the route splitting. At each price elasticity level, there is no systematic decline in the percentage of split traffic, even though link capacities are increased by more than 50 times. The result suggests that the ability to distribute a stream, possibly non-uniformly, to multiple paths is an important traffic engineering function that will not be obsolesced by having "fat" bandwidth pipes in future. However, at each capacity level, the split ratio decreases (non-uniformly, however) as price elasticity increases. The effect of price elasticity on route splitting is substantial, which can be explained by Theorems 3-2 and 3-3. High elasticity induces minimum-hop routing(Theorem 3-2), and leads to a larger proportion of direct traffic in total demand(Theorem 3-3). As a result, more demands will only have one minimum cost admissible route, so the proportion of traffic that can be split is reduced.

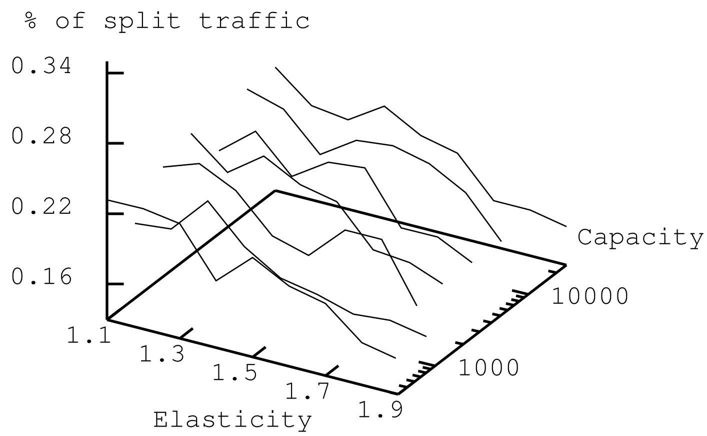

Figure 6. Percentage of Split Traffic

5. CONCLUSIONS

We have developed a generalized network engineering model to maximize total revenue in a multi-service network. For this model, we have obtained the optimal solution, and derived its salient properties. We have shown that link shadow costs provide the base for unifying pricing and routing decisions. The service with the highest price-elasticity grows faster than any other service when network capacity is scaled to increase uniformly across links. We have also shown that high price elasticity leads to direct-routed traffic dominating non-direct-routed traffic on links. We have demonstrated that neither uniform capacity expansion nor high price-elasticity can remove traffic-splitting from optimal routing. However, high price-elasticity induces minimum-hop routing and reduces the amount of split traffic.

Our joint optimization model can be connected to other decision problems related to providing network services. For example, we are considering using shadow costs, derived from this model, to regulate capacity provisioning.

6. APPENDIX

6.1. Proof of Lemma 1

Let $l_{min} = \arg\min_{l} \lambda_l$, and σ_{min} be the end node pair of l_{min}. Then for any stream (s, σ_{min}), l_{min} is the only minimum cost admissible route, so all demands $D_{s\sigma_{min}}$ are carried on l_{min}, i.e.:

$$M(s, \sigma_{min}) = \{l_{min}\},\ \mu_{s\sigma_{min}} = \lambda_{l_{min}},\ \text{and } D_{s\sigma_{min}} = X_{sl_{min}}$$

Apply (18) to the above and notice that $X_{sl_{min}} \leq C_{l_{min}}$ by the link capacity constraint:

$$\lambda_{l_{min}} = \mu_{s\sigma_{min}} = (\tilde{A}_{s\sigma_{min}}/D_{s\sigma_{min}})^{1/\epsilon_s} = (\tilde{A}_{s\sigma_{min}}/X_{sl_{min}})^{1/\epsilon_s} \geq (\tilde{A}_{s\sigma_{min}}/C_{l_{min}})^{1/\epsilon_s}, \text{ and} \tag{30}$$

$$(\tilde{A}_{s,min}/C_{max})^{1/\epsilon_s} \leq (\tilde{A}_{s\sigma_{min}}/C_{l_{min}})^{1/\epsilon_s} \leq \lambda_{l_{min}} \leq \lambda_l \text{ for all } l \tag{31}$$

As for the upper bound, let $l_{max} = \arg\max_{l} \lambda_l$, and consider all demands that has l_{max} on at least one of their minimum cost admissible route (i.e.: $\{(s', \sigma')|\exists r \in M(s', \sigma') : l \in r\}$. These are only demands that can be carried on l_{max}. By definition of (s', σ') and equation (12),

$$\sum_{s',\sigma'} D_{s'\sigma'} \geq \sum_{s',\sigma'} \sum_{r \in M(s',\sigma'):l_{max} \in r} X_{s'r} = C_{l_{max}} \tag{32}$$

$$\mu_{s'\sigma'} \geq \lambda_{l_{max}} \Rightarrow \sum_{s',\sigma'} D_{s',\sigma'} = \sum_{s',\sigma'} \tilde{A}_{s'\sigma'} \mu_{s'\sigma'}^{-\epsilon_{s'}} \leq \sum_{s',\sigma'} \tilde{A}_{s'\sigma'} \lambda_{l_{max}}^{-\epsilon_{s'}} \leq \sum_{s,\sigma} \tilde{A}_{s\sigma} \lambda_{l_{max}}^{-\epsilon_s} \tag{33}$$

The last inequality comes from that $\{(s',\sigma')\}$ is a subset of $\{(s,\sigma)\}$. It follows that:

$$C_{l_{max}} \leq \sum_{s,\sigma} \tilde{A}_{s\sigma} \lambda_{l_{max}}^{-\epsilon_s} \leq \max[\lambda_{l_{max}}^{-\epsilon_{min}}, \lambda_{l_{max}}^{-\epsilon_{max}}] \sum_{s,\sigma} \tilde{A}_{s\sigma} \tag{34}$$

$$\lambda_l \leq \lambda_{l_{max}} \leq \max[(\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min})^{1/\epsilon_{min}}, (\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min})^{1/\epsilon_{max}}] \quad \text{for all } l \tag{35}$$

6.2. Proof of Theorem 3-1

Since $\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min0} \leq 1$, $(\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min0})^{1/\epsilon_{min}} \leq (\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min0})^{1/\epsilon_{max}}$

On applying Lemma 1,

$$(\frac{\tilde{A}_{min}}{C_{max0}})^{1/\epsilon_{max}} \leq \lambda_l(1) \leq (\frac{\sum_{s,\sigma} \tilde{A}_{s\sigma}}{C_{min0}})^{1/\epsilon_{max}} \quad \text{and} \quad (\frac{\tilde{A}_{min}}{mC_{max0}})^{1/\epsilon_{max}} \leq \lambda_l(m) \leq (\frac{\sum_{s,\sigma} \tilde{A}_{s\sigma}}{mC_{min0}})^{1/\epsilon_{max}} \tag{36}$$

Therefore, $m^{-1/\epsilon_{max}}/G \leq \dfrac{\lambda_l(m)}{\lambda_l(1)} \leq G m^{-1/\epsilon_{max}}$ where: $G = (\dfrac{C_{max0} \sum_{s,\sigma} \tilde{A}_{s\sigma}}{C_{min0} \tilde{A}_{min}})^{1/\epsilon_{max}}$

6.3. Proof of Theorem 3-2

Assume a network has N nodes and L links. For each service s, there are at most $N^2 - L$ non-directly-routed streams. Let $D_{s\sigma}^t$ be any one of the non-directly-routed demand, and $r = (l_1, l_2, ..., l_n)$ be a route that carries $D_{s\sigma}^t$, then:

$$D_{s\sigma}^t = A_{s\sigma}^t [\frac{\epsilon_s}{\epsilon_s - 1}(\lambda_{l_1} + \lambda_{l_2} + \ldots + \lambda_{l_n})]^{-\epsilon_s} \tag{37}$$

Let $\bar{l} = \{\bar{l} \in r : \lambda_{\bar{l}} = \min_{l \in r} \lambda_l\}$, and $\bar{\sigma}$ be the end-point pair of $\bar{l}$, then the direct stream between $\bar{\sigma}$:

$$D_{s\bar{\sigma}}^d \geq A_{s\bar{\sigma}}^d (\frac{\epsilon_s}{\epsilon_s - 1} \lambda_{\bar{l}})^{-\epsilon_s} \tag{38}$$

so

$$\frac{D_{s\sigma}^t}{D_{s\bar{\sigma}}^d} \leq \frac{A_{s\sigma}^t}{A_{s\bar{\sigma}}^d} (\frac{\lambda_{l_1} + \ldots + \lambda_{l_n}}{\lambda_{\bar{l}}})^{-\epsilon_s} \leq 2^{-\epsilon_s} \frac{A_{s,max}}{A_{s,min}} \quad \text{and} \quad D_{s\sigma}^t \leq 2^{-\epsilon_s} \frac{A_{s,max}}{A_{s,min}} D_{s\bar{\sigma}}^d \leq 2^{-\epsilon_s} \frac{A_{s,max}}{A_{s,min}} D_s^d \tag{39}$$

Since the above holds for each of the $N^2 - L$ non-directly-routed streams,

$$\frac{D_s^t}{D_s^d} \leq 2^{-\epsilon_s} (N^2 - L) \frac{A_{s,max}}{A_{s,min}} \quad \text{where } D_s^t + D_s^d = \sum_\sigma D_{s\sigma}. \tag{40}$$

$$\lim_{\epsilon_s \to \infty} D_s^d / \sum_\sigma D_{s\sigma} = 1. \tag{41}$$

6.4. Proof of Theorem 3-3

Since $\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min0} \leq 1$, $\quad (\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min0})^{1/\epsilon_{min}} \leq (\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min0})^{1/\epsilon_{max}}$

On applying Lemma 1,

$$\begin{aligned} \sum_{l \in r_2} \lambda_l &\leq H(r_2)(\sum_{s,\sigma} \tilde{A}_{s\sigma}/C_{min})^{1/\epsilon_{max}} \leq H(r_2)(\sum_{s,\sigma} A_{s\sigma}/C_{min})^{1/\epsilon_{max}} \\ \sum_{l \in r_1} \lambda_l &\geq H(r_1)(\tilde{A}_{\bar{s},min}/C_{max})^{1/\epsilon_{max}} = H(r_1) \frac{(\epsilon_{max} - 1)}{\epsilon_{max}} (\frac{A_{\bar{s},max}}{C_{max}})^{1/\epsilon_{max}} \end{aligned} \tag{42}$$

where $\bar{s} = \{\bar{s} \in S : \epsilon_{\bar{s}} = \max_s \epsilon_s\}$

$$H(r_1) > H(r_2) \quad \Rightarrow \quad \lim_{\epsilon_{max} \to +\infty} \frac{\sum_{l \in r1} \lambda_l}{\sum_{l \in r2} \lambda_l} = \frac{H(r_1)}{H(r_2)} \lim_{\epsilon_{max} \to +\infty} \frac{(\epsilon_{max} - 1)}{\epsilon_{max}} (\frac{C_{min} A_{\bar{s},max}}{C_{max} \sum_{s,\sigma} A_{s\sigma}})^{1/\epsilon_{max}} > 1 \tag{43}$$

REFERENCES

1. Ahuja, R. K., Magnanti, T. L., and J. B. Orlin, *Network Flows.* Prentice-Hall, Inc., Upper Saddle River, NJ, 1993.
2. P. Aukia *et al.*, RATES: A Server for MPLS Traffic Engineering, *IEEE Network*, **14**, pp. 34-41, March/April 2000.
3. Awduche, D. O., Chiu, A., Elwalid, A., Widjaja, I., and X. Xiao, A Framework for Internet Traffic Engineering, *Internet Draft, IETF*, July 2001.
4. Blake, S., Black, D., Carlson, M., Davis, D., Wang, Z., and W. Weiss, An Architecture for Differentiated Services, *RFC2475, IETF.*
5. Courcoubetis, C. A., Dimakis, A., and M. I. Reiman, Providing Bandwidth Guarantees Over a Best-effort Network: Call Admission and Pricing, *Proceedings of IEEE INFOCOM2001*, pp. 459-467, April 2001.
6. Dixit, A. K., *Optimization in Economics Theory*, 2nd edition, Oxford University Press, Inc., New York, 1990.
7. A. Feldmann *et al.*, NetScope: Traffic Engineering for IP Networks, *IEEE Network*, **14**, pp. 11-19, March/April 2001.
8. Fortz, B. and M. Thorup, Internet Traffic Engineering by Optimizing OSPF Weights, *Proceedings of IEEE INFOCOM2000*, pp. 519-528, 2000.
9. Intriligator, M. D., *Mathematical Optimization and Economic Theory*, Englewood Cliffs, Prentice-Hall, NJ, 1971.
10. Kelly, F., Maulloo, A., and D. Tan, Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability, *Journal of the Operational Research Society*, **49**, pp. 237-252, 1998
11. Keon, N. and G. Anandalingam, A New Pricing Model for Competitive Telecommunications Services Using Congestion Discounts, *pre-print*, 2001.
12. Pindyck, R. S. and D. L. Rubinfeld, *Microeconomics*, 2nd edition, Macmillan Publishing Company, New York, 1992.
13. Lagoa, C. and H. Che, Decentralized Optimal Traffic Engineering in the Internet, *Computer Communications Review*, **30**, pp. 39-47, October, 2000.
14. Lanning, S. G., Mitra, D., Wang, Q., and M. H. Wright, Optimal Planning for Optical Transport Networks, *Phil.Trans.Royal Soc.London A*, **358**, pp. 2183-2196, August 2000.
15. Lanning, S. G., Neuman, W. R., and S. O'Donnell, A Taxonomy of Communications Demand, *Proceedings of 27th Telecommunications Policy Research Conference*, Alexandria, VA, September 1999.
16. MacKie-Mason, J. K. and H. Varian, Pricing Congestible Resources," *IEEE Journal of Selected Areas in Communications*, **13**, pp. 1141-1149, September 1995.
17. Mitra, D. and K. G. Ramakrishnan, A Case Study of Multiservice Multipriority Traffic Engineering Design for Data Networks, *Proceedings of IEEE GLOBECOM 99*, pp. 1077-1083, December 1999.
18. Mitra, D., Ramakrishnan, K. G., and Q. Wang, Combined Economic Modeling and Traffic Engineering: Joint Optimization of Pricing and Routing in Multi-Service Networks, *Bell Labs Technical Memorandum*, 2001.
19. Moy, J. T., *OSPF Anatomy of an Internet Routing Protocol*, Addison-Wesley, Reading, MA, 1998.
20. P. Trimintzios *et al.* A Management and Control Architecture for Providing IP Differentiated Services in MPLS-Based Networks, *IEEE Communications*, **39**, pp. 80-87, May 2001.
21. Vavasis, A., *Nonlinear Optimization: Complexity Issues*, Oxford University Press, New York, 1991.
22. Wang, Q., Peha, J. M., and M. A. Sirbu, Optimal Pricing for Integrated Services Networks, *Internet Economics*, edited by L. W. McKnight and J. P. Bailey, pp. 352-376, MIT Press, Cambridge, MA, 1997.