The Pricing of the Grade of Service Guarantees in the Service Overlay Networks

Ngok Lam, Zbigniew Dziong, and Lorne G. Mason

***Abstract*—** We studied a class of Service Overlay Network (SON) capacity allocation problem. By analyzing the problem with two different nonlinear optimization formulations, we show that the prices of offering service guarantees are closely related to a set of Lagrange multipliers. Moreover, if the Grade of service (GoS) constraints are not hard requirements, the network design resulting from the set of prices is on the Pareto frontier of a bi-objective optimization problem. A scheme was developed to derive the prices for various classes of customers by referring to the Lagrange multipliers. The major contribution of the article is the use of the Lagrange multipliers to provide a set of Pareto efficient prices in providing GoS guarantees.

***Keywords*—**Service Guarantees, Network Pricing, Network Management.

I. Introduction

The demand for end-to-end Quality of Service (QoS) guarantees in the Internet has increased significantly due to the introduction of new applications like VoIP, online gaming, and video conferencing. This poses a major challenge to the current Internet architecture. Owing to historical reasons, the Internet consists of a large collection of independent Autonomous Systems (ASes). In order to ensure end-to-end QoS guarantees of the data, one has to build a multi-lateral business relationship with all the independent ASes his data transit. This makes it unrealistic to obtain end-to-end QoS guarantees. A higher level mechanism on the top of the Internet known as Service Overlay Network (SON) is thus proposed to alleviate this problem [10]. The SON network operates in a manner similar to a virtual network. The SON operator owns the SON gateways which are placed in strategic locations. To realize the SON network, the SON operator leases bandwidths with QoS guarantees from the underlying Autonomous Systems, (ASes) in the form of Service Level Agreements (SLAs). The leased bandwidths act as logical links that connect the SON gateways. Once all the logical links are in place, the SON is realized and the overlay network formed is under the administration of a single authority. Because the SON is administrated by a single operator, it is capable of providing end-to-end QoS guarantees for the value-added services provided by it. A user with access to the Internet can access the service gateway to use the value-added services, provided the hosts holding the contents are also connected to some SON service gateways. In a SON, the connections are classified by the origin and the destination (OD) gateways. Users pay the service charge based on the origins and destinations of their connections as well as their connection durations. Figure 1 shows an example of the SON network.

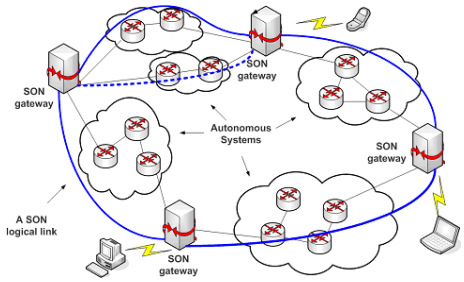

Fig. 1. An example SON network.

Once the SON network has been designed and realized, a major challenge to the SON operator would be to charge the services appropriately. The prices should generate maximal economic benefits to the operator. Yet they should also be reasonable with respect to the users' budgets. In this article, we introduce a set of pricing metric that enables the SON network to generate profit optimally while providing the Grade of Service guarantees (GoS) to the users. It can be shown this pricing metric is minimal, and it is a Pareto efficient solution to a bi-objective optimization problem that maximizes the utilities of both the operator and the user. This article is organized as follows: Section II is the description of the problem assumptions and formulations, Section III discusses the major results, Section IV shows a simple example that illustrates the results, Section V is the conclusion section that concludes the results obtained.

II. Problem Formulations

A. The optimization models

To decide the optimal amount of bandwidths to be allocated on the logical links, operator usually resort to two distinct yet related mathematical models, namely the Maximum Profit(MP) and the Minimum Cost(MC) models. We assume that the operator considers profit as the primary performance measure for the network. Therefore the operator is assumed to employ the MP model. By considering the SON as a loss network [5],

Ngok Lam is with the Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7 (e-mail: ngok.lam@mcgill.ca).

Zbigniew Dziong is with Department of Electrical Engineering, Ecole de Technologie Superieure, 1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1k3 (e-mail: zdziong@ele.etsmtl.ca).

Lorne G. Mason is with the Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7 (e-mail: lorne.mason@mail.mcgill.ca).

the MP model is given by (2.1a). The problem is assumed to be solved using the Lagrangian relaxation approach in [3]. The first order optimality condition of (2.1a) is listed in (2.1b). We shall denote formulation (2.1a) as "F1" in the rest of this article.

$$\max_{N_s} \quad \sum_{ij} \lambda^{ij} w^{ij}(1-B^{ij}) - \sum_{s} C_s(N_s)$$

$$N_s \geq 0 \qquad z_s \qquad (2.1a)$$

$$c_s = \sum_{ij} \lambda^{ij} w^{ij} \left(\frac{-\partial B^{ij}}{\partial N_s^*}\right) \quad \forall s, \text{ s.t. } N_s^* > 0 \qquad (2.1b)$$

In the formulation (2.1a), λ^{ij} is the given poissonian connection arrival intensity demanding the connection of the node pair *(i,j)* (i.e. origin gateway is *i*, destination gateway is *j*), w^{ij} is the expected service charges (prices) paid by an admitted *(i,j)* connection. The symbol B^{ij} denotes the analytical end-to-end blocking probability for connections of the node pair *(i,j)*, due to lack of available resource. It is an end-to-end blocking probability dependent on the (optimal) routing scheme employed. The capacity of a link *s* is denoted by N_s and it is a decision variable of this problem. The function $C_s(.)$ is the cost function that quantifies the cost rate of allocating N_s units of capacities on link *s* (based on some SLA) and it is assumed to be a linear function of the variable N_s. The variables z_s is the Lagrange multipliers to ensure non-negative capacity assignments.

The users of the SON may desire a certain level of service guarantee so that their connection requests are granted with probabilities higher than some thresholds. If the operator is to fulfill this expectation, they need to allocate additional resources on the logical links. This introduces extra costs. The minimum cost design that satisfies the grade of service expectations is a solution which requires the minimum investment to realize the service guarantees. It is the solution of the formulation (2.2a). The corresponding first order optimality condition is given by (2.2b)

$$\min_{N_s} \quad \sum_{s} C_s(N_s)$$

$$B^{ij} \leq L^{ij} \qquad v_{ij}$$

$$N_s \geq 0 \qquad z_s \qquad (2.2a)$$

$$c_s = \sum_{ij} v_{ij}^* \left(\frac{-\partial B^{ij}}{\partial N_s^*}\right) \qquad \forall s, \text{ s.t. } N_s^* > 0 \qquad (2.2b)$$

Two new notations are being introduced in (2.2). The first new notation is L^{ij}, which specifies the user desired threshold on the end-to-end blocking probability for connections of the OD pair (*i,j*). The second new notation is v_{ij}, it is the Lagrange multiplier corresponds to the GoS constraint. We shall denote formulation (2.2a) as "F2" in the remainder of the article. Without loss of generality, we assumed that both the F1 and F2 formulations employ the same (optimal) routing scheme in the routing layer. We shall show in the following sections that the multipliers v_{ij} from F2 is a set of Pareto efficient solution that maximizes the user utility and the objective of F1.

B. The end-to-end blocking probabilities

The end-to-end blocking function B^{ij} is a fundamental component of the formulations F1 and F2. The actual functional form of B^{ij} varies with routing schemes [4]. We take a different perspective and derive it by using the insight that the blocking function can be approximated based on the link connection intensities (at equilibrium) and the capacities [4]. Techniques from the reliability theory [7] were employed to devise the general functional form for B^{ij}, regardless of the actual routing scheme employed. The B^{ij} function obtained below is based on the reduced load approximation model [9], which assumed statistical link independence and Poisson link arrival rates.

We consider the *collection* of network paths, that connect a particular origin node *i* with a particular destination node *j*, as a complete system. The task of this system is to serve the connections between the node pair (*i,j*). Assume the network links are independent of one another. The links in the *collection* of paths are the independent components of the system. Denote these links by *s* and let R_{ij} be a set that contains all these links. Define an indicator variable y_s for the link *s*, whereas y_s equals to *zero* if link *s* has enough resource to admit at least one connection, and y_s equals to *one* if link *s* does not have resource to serve any connection. The expected value of y_s is therefore the blocking probability of link *s*. According to the reliability theory [7], a Boolean function $\phi(Y)$ that indicates whether the system has the available resource for new (*i,j*) connections can be defined by taking Y=[y_s] as the input. The complement of it, $\bar{\phi}(Y) = 1\text{-}\phi(Y)$ is another Boolean function that indicates whether the system has ran out of resource for new (*i,j*) connections. Thus the expected value of $\bar{\phi}(Y)$ is the end-to-end blocking probability for connection pair (*i,j*). Since y_s are independent zero-one random variables and $\bar{\phi}(Y)$ is a Boolean function, we can perform the Shannon decomposition on the function $\bar{\phi}(Y)$. By using an arbitrary link *s* as the pivot we have expression (2.3).

$$\begin{aligned}\bar{\phi}(Y) &= y_s \bar{\phi}(1_s, Y) + (1 - y_s)\bar{\phi}(0_s, Y) \\ &= \bar{\phi}(0_s, Y) + [\bar{\phi}(1_s, Y) - \bar{\phi}(0_s, Y)]y_s \end{aligned} \qquad (2.3)$$

Where $(0_s, Y)$ and $(1_s, Y)$ are the status vectors that differs only in the s^{th} link. The functions $\bar{\phi}(0_s, Y)$ and $\bar{\phi}(1_s, Y)$ indicates that whether the system has been blocked given that the link *s* is in admissible status/has been blocked. By the definition of y_s, the expectation E[y_s] is the blocking probabilities of link *s*. Assume the links are independent and link arrival rates are Poisson, we have expression (2.4). The expectations E[y_s] and E[$y_{s'\neq s}$] are replaced by the Erlang-B Loss formula $E_s(.)$ and the vector $E_{s'\neq s}$ respectively in (2.4). The vector $E_{s'\neq s}$ denotes the collection of Erlang-B loss functions for all the links *s'* such that *s'*≠*s*. The continuous extension of Erlang-B formula suggested in [1] is being used throughout this article and it is shown in (2.5).

$$E[\bar{\phi}(Y)] = f_{1,s}^{ij}(E_{s'\neq s}) + f_{2,s}^{ij}(E_{s'\neq s})E_s(.) \qquad (2.4)$$

It should be clear now that B^{ij}=E[$\bar{\phi}(Y)$] is a reduced-load approximation of the end-to-end blocking probability, as link independence and Poisson link arrival rates are assumed. Note that $f_{2,s}^{ij} \geq 0$, if the end-to-end blocking probability B^{ij} is strictly decreasing in the presence of additional available link ($f_{2,s}^{ij}$ is the Birnbaum's importance measure of link s in the context of

reliability theory). This is a monotonic property we imposed on the routing scheme and it is assumed throughout the article. We also assume another monotonic property such that the routing scheme does not decrease the (equilibrium) link connection intensity as the capacity of the link increases. Finally we assume that the routing objective function is uni-modal with respect to the capacities.

$$E_s(\lambda_s, N_s) = \{\lambda_s \int_0^{+\infty} e^{-\lambda_s z}(1+z)^{N_s} dz\}^{-1} \quad (2.5)$$

Since the value of the Erlang-B formula can be uniquely determined by the link capacity and the link connection arrival rate [1], therefore (2.4) is rewritten to (2.6) to explicitly state the dependence of B^{ij}'s on the (equilibrium) link connection intensities and link capacities.

$$B^{ij} = f_{1,s}^{ij}(\lambda_{s'\in R_{ij}, s'\neq s}, N_{s'\in R_{ij}, s'\neq s}) + f_{2,s}^{ij}(\lambda_{s'\in R_{ij}, s'\neq s}, N_{s'\in R_{ij}, s'\neq s}) E_s(\lambda_s, N_s) \quad (2.6)$$

Expression (2.6) is valid for all the link s. So we can represent B^{ij} in the form of $f_{1,s}^{ij} + f_{2,s}^{ij} E_s(\lambda_s, N_s)$ for every link s, where $f_{1,s}^{ij}$ and $f_{2,s}^{ij}$ are independent of the link s. Note that $\lambda^{ij} f_{2,s}^{ij}$ is the connection intensities to the link s from the OD pair (i,j).

III. Main results

A. Optimal Grade of Service Guarantees

We shall show in this section that, if the service charge is high, and if the operator's objective is to maximize the profit (using F1), then the optimal decision for the operator is to offer better GoS guarantees. Intuitively this means that the operator should deliver a lower blocking probability to high-reward connections so as not to miss profit making opportunities.

We assume the relaxation scheme in [3] is being employed to solve F1. This approach solves the *exact* first order condition instead of the linearized approximation (i.e the Netwon's method [6]) in each iteration. The scheme solves the set of first order optimality conditions based on the previous solution, and the iteration continues until a stationary point is reached.

Lemma 1 below establishes the relation between the optimal capacities allocated and the magnitude of service charges. Consider equation (3.1), where c_s is a positive constant, v is a positive number, $E_s(.)$ is the Erlang-B formula as defined in (2.5), λ_s is the connection intensity on link s, N_s^v is the capacity allocated on link s. Then the following lemma holds.

Lemma 1: If λ_s is fixed in (3.1), and v^1, v^2 are two real numbers, where $v^1 > v^2 > 0$, then we have $N_s^{v^1} > N_s^{v^2}$, where $N_s^{v^1}$ and $N_s^{v^2}$ are the values of N_s^v in (3.1) And v^1, v^2 are the values of v in (3.1).

$$c_s = v(-\frac{\partial E_s(\lambda_s, N_s^v)}{\partial N_s^v}) \quad (3.1)$$

Proof:

It is known that the Erlang-B formula is a C^∞ function [2], which is strictly convex in the capacity [1]. Therefore for a fixed λ_s, the function $\frac{-\partial E(\lambda_s, N_s^v)}{\partial N_s^v}$ is strictly decreasing and continuous. So the expression $v(-\frac{\partial E_s(\lambda_s, N_s^v)}{\partial N_s^v})$ is continuous and strictly decreasing in N_s^v, where v is a positive number. As a result for a constant c_s, the larger the value v, the smaller the expression $(\frac{-\partial E(\lambda_s, N_s)}{\partial N_s^v})$ will be required to satisfy the equality condition of expression (3.1). This therefore requires a larger N_s^v value. As a result if $v^1 > v^2 > 0$ and if $N_s^{v^1}$ and $N_s^{v^2}$ both exist then $N_s^{v^1} > N_s^{v^2}$. □

Theorem 1: Assume the routing scheme does not decrease the (equilibrium) link connection intensity as the capacity of a link increases. Then the optimal GoS derived by F1 is an increasing function of the service charge vector $\boldsymbol{W}=[w^{ij}]$.

Proof:

Suppose the optimization approach in [3] is employed to solve the problem. Assume the method converges to the optimal solution. By using (2.5), the first order optimality condition (2.1a) is re-written to (3.2), note that (3.2) represents n set of equations where n equals to the number of links in the network.

$$c_s = [\sum_{ij}(\lambda^{ij} w^{ij}) \times f_2^{ij}(\lambda_{s'\in R_{ij}, s'\neq s}, N_{s'\in R_{ij}, s'\neq s})](\frac{-\partial E_s(\lambda_s, N_s)}{\partial N_s}) \quad (3.2)$$

Now consider two vectors of service charges, $\boldsymbol{W}=[w^{ij}]$ and $\boldsymbol{W'}=[w^{ij} + \Delta w^{ij}]$, where $\Delta w^{ij} > 0$. Assume that the optimal solution with respect to the vector $\boldsymbol{W}$ is denoted by the tuple $(\Lambda^W, N^W(\Lambda^W))$, where $\Lambda^W = [\lambda_s^W]$ is the link connection intensity vector decided by some optimal routing rules, $N^W(\Lambda^W) = [N_s]$ is the optimal capacity allocation on the logical links. Now consider the case that $(\Lambda^W, N^W(\Lambda^W))$ is regarded as the initial solution of the F1 (with parameters $\boldsymbol{W'}$). Substitute $(\Lambda^W, N^W(\Lambda^W))$ into the optimality condition (3.2), note that the service charges are now $\boldsymbol{W'}$, and we have $\sum_{ij}(\lambda^{ij} w^{ij} + \lambda^{ij}\Delta w^{ij}) \times f_2^{ij}(\lambda_{s'\in R_{ij}, s'\neq s}, N_{s'\in R_{ij}, s'\neq s}) > \sum_{ij}(\lambda^{ij} w^{ij}) \times f_2^{ij}(\lambda_{s'\in R_{ij}, s'\neq s}, N_{s'\in R_{ij}, s'\neq s})$, by lemma (1) the allocated capacities on all the link strictly increases at the end of the first iteration. Since the equilibrium connection intensity is non-decreasing when link capacity increases, therefore and $f_{2,s}^{ij}$ increases at the end of the first iteration. This makes $\sum_{ij}(\lambda^{ij} w^{ij} + \lambda^{ij}\Delta w^{ij}) \times f_2^{ij}(\lambda_{s'\in R_{ij}, s'\neq s}, N_{s'\in R_{ij}, s'\neq s})$ increases further, and the capacities are augmented further. This augmentation process continues until the optimal solution $N^{W'}(\Lambda^{W'})$ is reached, and the system of equations in (3.2) reach a fixed point. Therefore we have $N^{W'}(\Lambda^{W'}) > N^W(\Lambda^W)$ at the optimality. Now because of the monotonic assumption of the routing scheme and also because $N^{W'}(\Lambda^{W'}) > N^W(\Lambda^W)$, the GoS guarantees offered by $N^{W'}(\Lambda^{W'})$ is strictly better than that being offered by $N^W(\Lambda^W)$. One can consider the set of resulting GoS guarantees as the optimal GoS guarantees, as they are achieved when the profit from the SON is maximized. □

Theorem 2: Assume the user-desired GoS guarantees are denoted by L^{ij}, (i.e. the users of OD pair (i,j) desires an end-to-end blocking probability of B^{ij} less than L^{ij}). Then the minimum charge they need to pay to the operator is defined by

(3.3). The symbols v_{ij}^* in (3.3) are the Lagrange multipliers of the formulation F2 with the set of user desired L^{ij} as constraints.

$$w^{ij} = v_{ij}^* / \lambda^{ij} \tag{3.3}$$

Proof:

The minimum cost assignment which satisfies the desired GoS levels satisfy the equations in (2.2b). Assume the operator design the SON network by using formulation F1. If the service charges defined in (3.3) are substituted into (2.1b), it is easy to see that expressions (2.2b) and (2.1b) become identical, the second order optimality conditions will also be the same (see [8] for details), and F1 gives the same capacity assignments as F2. Therefore if the users pay the service charges defined by (3.3), they will get the desired GoS guarantees from the operator even if the operator designs the networking using F1. Theorem 1 implies the GoS level decreases when the service charge decreases. Assume the same conditions on the routing scheme as being assumed in theorem 1 hold, then the service charges defined by (3.3) is the minimum charge the users need to pay in order to enjoy the desired levels of GoS guarantee if the operator designs the network using F1. □

Theorems 1 and 2 can be illustrated through the use of a figure below. Figure 2 shows the profit contours for a one-link network, there is only one OD pair (*i,j*), and it is connected together by a link. Each blue line in figure 2 corresponds to the expected profit from the link under a particular connection service charge w^{ij}. The *x*-axis corresponds to the capacity assigned to this link and the *y*-axis corresponds to the expected profit rate from the link. The red dots are the points that generate the maximum expected profits with respect to the w^{ij}, thus the red dots denote the optimal solutions of F1 under the parameters w^{ij}. For clear illustration, an arrow is drawn to point the direction of increasing w^{ij}. As w^{ij} increases, the optimal capacity allocation (i.e. the *x* values of the red points) also increases. A vertical black line was drawn to indicate the minimum capacity required for a desired level of GoS. There is a red dot that intersects with the black line. This particular red dot is the optimal solution of the formulation F2 with the GoS constraints (it is also a solution of F1 with respect to the particular parameter w^{ij}). The reward w^{ij} on this very red dot, corresponds to the value of $w^{ij} = v_{ij}^* / \lambda^{ij}$ as defined in (3.3). This is the minimum service charge the users need to pay in order to enjoy the desired GoS level. If the service charge is higher than this particular w^{ij}, the connections will be assigned larger capacities and enjoy even better GoS guarantees. The service charges defined in (3.3) can be interpreted as the minimum service charges that drive the network operator to offer the levels of GoS desired by its users. We shall show in the following sub-section that this set of service charges is a Pareto efficient solution to a bi-objective optimization problem.

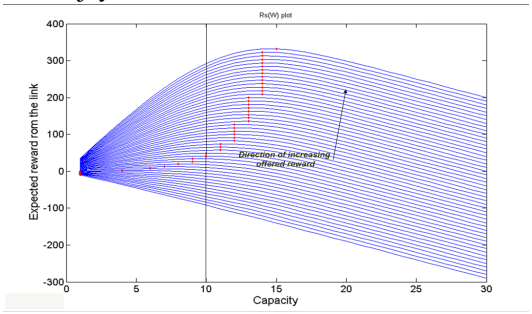

Fig. 2. An example to illustrate that GoS offered by F1 is an increasing function of the service charges

B. *Pareto efficient pricing*

Assume that the utility of operator is an increasing function of total expected profit gained from the SON, then F1 will be employed to design the SON network. Denote the utility function of the users of OD pair (*i,j*) by U_{ij}. Assume all the user of the connection pair (*i,j*) desire a certain level of GoS guarantee, denote it by L^{ij}. If this level of GoS guarantee is not achieved, then we have $U_{ij}(L^{ij}, M^{ij}) > U_{ij}(\overline{L^{ij}}, x)$, for all and $0 \le x < M^{ij}$, where $\overline{L^{ij}}$ is an abuse of the symbol to indicate that the GoS level is below L^{ij}, and M^{ij} is the maximum amount of money that users of the OD pair (*i,j*) are willing to pay for the service. Assume that the users always prefer low service charge so we have $U_{ij}(B^{ij}, x^{ij}) > U_{ij}(B^{ij}, y^{ij})$, $M^{ij} \ge y^{ij} > x^{ij}$. Where B^{ij} is the GoS perceived by the users, x^{ij} and y^{ij} denote the monetary values the users pay. Assume further that the utility U_{ij} only depends on the service charge when L^{ij} is satisfied. Consider the problem of maximizing both the user utility and the operator utility in the bi-objective optimization formulation as shown in (3.4), assume this problem is feasible.

$$\begin{aligned} &\max_{w^{ij}} f_o = \max_{N_s} \sum_{ij} \lambda^{ij} w^{ij} (1 - B^{ij}) - \sum_{s} C_s(N_s) \\ &\max_{w^{ij}} f_c = \sum_{ij} U_{ij}(B^{ij}, w^{ij}) \\ &s.t. \quad 0 \le w^{ij} \le M^{ij} \quad \forall ij \end{aligned} \tag{3.4}$$

Lemma 2: f_o is an increasing function of w^{ij}.

Proof:

Consider the case that w^{ij} is increased by $\Delta w^{ij}>0$, assume the original service charge vector be $\boldsymbol{W}=[w^{ij}]$, denote a new reward vector by $\boldsymbol{W'}$ where $\boldsymbol{W'}$ is larger than $\boldsymbol{W}$ by Δw^{ij} at the ij^{th} element. Substitute $\boldsymbol{W'}$ into f_o at the optimal solution of the original $\boldsymbol{W}$ (i.e. B^{ij} and N_s remain unchanged), the value increases even without re-optimization. Since the value of f_o can only increase after re-optimization, so we have f_o ($\boldsymbol{W'}$)> f_o ($\boldsymbol{W}$). □

It can be shown that the set of service charges defined in (3.3) is a minimum Pareto efficient solution to the bi-objective problem in (3.4).

Theorem 3. The set of service charges defined in (3.3) is the minimum Pareto efficient solution to (3.4).

Proof:

Since w^{ij} in (3.4) is bounded and closed, the feasible set of (3.4) is compact. Now consider two possible deviations of w^{ij}.

a) if w^{ij} is increased by a positive amount of Δw^{ij}, assume this

move is feasible, then f_o increases according to lemma (2). Now since **W'>W,** so according to theorem 1 and 2, the preferred GoS levels are all satisfied. From the definition of U_{ij} we know that $U_{ij}(w^{ij}) > U_{ij}(w^{ij} + \Delta w^{ij})$. So f_o is improved by increasing w^{ij} (by Lemma 2) but f_c is worsened, and **W** is not dominated by **W'**.

b) if w^{ij} is decreased by a positive amount of Δw^{ij}, then according to lemma (2), f_o decreases. Moreover since **W'<W**, then according to theorems 1 and 2, the GoS desired by the ij^{th} users is not satisfied. From the definition of user utility we have $U_{ij}(L^{ij}, w^{ij}) \geq U_{ij}(L^{ij}, M^{ij}) > U_{ij}(\overline{L^{ij}}, w^{ij} - \Delta w^{ij})$. Therefore both f_o and f_c are worsened, and ***W***' is dominated by ***W***.

From part *b* of theorem 3, It can be seen that ***W*** is a minimum Pareto efficient vector. □

IV. A SIMPLE EXAMPLE

We shall show the results using an example. Consider a simple SON network in figure 3. Assume there are three Poisson streams of connections, with intensities λ_{AB}=10 units per unit time, λ_{CB} =15 units per unit time, λ_{AC} =20 units per unit time respectively. To make the discussion simple, all the connection streams are routed through the direct links. The mean holding times of the connections are assumed to be identically distributed with unit mean. The costs of leasing one unit of bandwidth for one unit of time are 5 units, 6 units and 7 units respective for links AB, CB, and AC, the allocated capacities are assumed to be integral values. Assume all the users desire a GoS level of 0.1. Assume that the operator considers profit and the primary performance metric of the network and F1 is

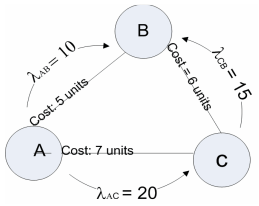

Figure 3. A simple SON network

employed to design the network. By using F2, we found the multipliers v^*_{AB}, v^*_{CB} and v^*_{AC} to be 182, 267 and 372 respectively (which translates to service charges of 18.2, 17.8 and 18.6 according to (3.3)). These service charges are substituted into F1 and the optimal solution is shown in table 1.

Table 1. Capacity allocation results for low service charge

Service charges (18.2, 17.8,18.6)	**Formulation F1**
GoS (λ_{AB}, λ_{CB},λ_{AC})	(0.084, 0.086, 0.085)
Allocated capacities on links (AB, CB ,AC)	(13, 18, 23)
Cost	334
Objective value	-417.13
Expected Profit rate	417.13

It can be seen from table 1 that all the desired GoS levels are achieved when the users pay according to the expression (3.3). The granularities of the GoS levels are not fine because of the integral nature of allocated capacities. But it nevertheless shows that the set of prices can indeed drive the operator to offer the desired level of GoS guarantees even though he is not obligated to do so.

V. CONCLUSIONS

We studied a class of Service Overlay Network (SON) capacity allocation problem. By assuming the profit as primary performance metric that the SON network operator is interested in, we derived a set of Pareto efficient service charges for the SON network. The service charges are derived from the set of multipliers v^*_{ij}, which are well known metrics that quantify the prices of the GoS constraints: the cost objective in (2.2a) can be improved by v^*_{ij} units if the corresponding GoS constraint is relaxed by one unit. So, intuitively this is also the amount of reward the users of the OD pair *(i,j)* should bring to the network so as to enjoy the GoS. To apply the results to a real SON business, the operator of SON can design the network using formulation F2, and charge according to the prices in (3.3). This set of prices makes the solutions of F1 and F2 coincide, which implies that the design the operator gets from F2 is a *maximum profit* design, but at the same time this design also gives the users a maximum utility. In this way the operator effectively delivers a network that maximizes both his profit and also the user utility (thus creating a win-win scenario). The study in the article shows that the Lagrange multipliers are capable of providing important pricing information to both the operator and the users. This additional set of pricing information is almost free of charge. First, it is because that the Lagrange multipliers are frequently the "side-products" of various numerical methods for solving an optimization problem (i.e. F2). Second, even if the values of the multipliers are not explicitly obtained, they could be computed relatively easy by solving a set of linear equations at the optimality (given that certain regularity condition holds).

REFERENCES

[1] A. Jagers and E. V. Doom, 1986. On the continued Erlang loss function. *Operations Research Letters*, Vol. 5, No. 1, pp. 43-46.

[2] J. S. Esteves, J. Craveirinha, and D. M. Cardoso, 2006. Second order conditions on the overflow traffic from the Erlang-B system, *Cadernos de Matemática, Universidade de Aveiro*, CM06/I-20.

[3] A. Girard and B. Liau, 1993. Dimensioning of adaptively routed networks, *IEEE transactions on networking*, Vol. 1, No. 4, pp. 460-468.

[4] A. Girard, 1993. Revenue optimization of telecommunication networks, *IEEE transactions on communications*, Vol. 41, No. 4, pp. 583-591.

[5] F.P. Kelly, 1988. Routing in circuit-switched networks: Optimization, Shadow Prices and Decentralization, *Advanced in applied probability*, Vol. 20, No. 1, pp. 112-144.

[6] D. P. Bertsekas, 1999. *Nonlinear Programming*, 2nd edition, Athena Scientific, USA.

[7] Wallace R. Blischke, D.N. Prabhakar Murthy, 2000. *Reliability Modeling, Prediction, and Optimization*. John Wiley & Sons, USA.

[8] Ngok Lam, Zbigniew Dziong, Lorne G. Mason, "The Equivalence of Maximum Profit and Minimum Cost Objectives in Service Overlay Network Design", Proceedings of the IADIS International Conference Telecommunications, Networks and Systems 2009, Algarve, Portugal, to appear..

[9] W. Whitt, 1985. Blocking when Service is Required from Several Facilities Simultaneously, *AT&T Technical Journal*, vol. 64, No. 8, pp. 1807-1856.

[10] Z. Duan, Z. L. Zhang, and Y. T. Hou, 2003. Service Overlay Networks: SLAs, QoS, and bandwidth provisioning, *IEEE/ACM Transactions on networking*, Vol.11, No. 6, pp. 870-883.