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

Ngok Lam, Lorne G. Mason
Department of Electrical and Computer Engineering,
McGill University, 3480 University Street, Montreal,
Quebec, Canada H3A 2A7
{ngok.lam,lorne.mason}@mcgill.ca

Zbigniew Dziong
Department of Electrical Engineering,
Ecole de Technologie Superieure,
Montreal, Canada
zdziong@ele.etsmtl.ca

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 multi-criteria optimization problem. The Pareto efficient prices for various classes of customers can therefore be derived by referring to the Lagrange multipliers.

Keywords: Service Guarantees, Network Pricing

I. INTRODUCTION

The Internet was designed to enable the free flow of data from one end point to another even when part of the network has failed. It was, however, not built with providing end-to-end Quality of Service (QoS) guarantees in mind. The connectivity of today's Internet is maintained by a large collection of independent Autonomous Systems (ASes). When data is transmitted from a source to a destination in the Internet, unless the source and destination are on the same network, the data has to traverse multiple independent ASes. In order to ensure end-to-end QoS guarantee of the data, the source needs to identify correctly all the independent ASes the data will transit, and then the source needs to request appropriate QoS guarantees from these ASes. The source might need doing this every time for new data transmissions. This unfortunate service-scape makes it highly unrealistic for a user to obtain end-to-end QoS guarantees in the Internet.

As the Internet is quickly turning into a multi-media platform for live and real-time contents, the demand for end-to-end Quality of Service (QoS) guarantees has increased significantly. Yet providing end-to-end QoS support on the decentralized Internet is nontrivial. This calls for the introduction of new networking concepts. Service overlay network (SON) is one such concepts[10]. It has recently gained significant attention in the networking community as a practical alternative to overcome some of the fundamental limitations of the Internet. The SON is capable of providing end-to-end QoS guarantees without the need of changing the Internet physical structure. It operates in a way similar to the virtual private network. An example SON is shown in figure 1.

Fig. 1. An example SON network.

The SON has two major building blocks, the gateways that are represented by the rectangular blocks in the figure, and the logical links that are represented by the blue lines. Each logical link is assumed to be provided by exactly one AS. Like the Internet routers, the SON gateways could be multi-homed, thus multiple ASes could be used to provide connectivity for the same logical link. This is illustrated by the upper left logical link in the figure (i.e. the dotted blue line represents an alternative AS for providing the same logical link). Since the SON is administrated by a single network operator, it is capable of providing QoS guarantees in an end-to-end sense on the top of the decentralized Internet. End-users connected to the Internet can access the service gateways to enjoy the QoS guaranteed communications.

A major challenge for the SON operator, after the SON network has been designed and realized, is 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' budget. 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 multi-criteria 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 performance measure for the network (i.e. the operator employs the MP model). By considering the SON as a loss network [5], 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 this article.

\[
\min_{s_i} \sum_j \lambda^j w^j (1-B^j) - \sum_s C_s(N_s)
\]

\[\lambda^j, N_s \geq 0 \quad \forall_s \quad z_s \geq 0 \quad \forall_s \]

In the formulation (2.1a), \(\lambda^j\) 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), \(w^j\) is the expected service charges (prices) paid by an admitted connection. The symbol \(B^j\) 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 underlying (optimal) routing scheme. 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 have some grade of QoS-guaranteed communication such that their requests for the QoS-guaranteed communication are granted with probabilities higher than some thresholds. If the operator is to fulfill these expectations, they need to allocate additional resources on the logical links. This introduces extra costs.

The minimum cost design that satisfies the GoS expectations is a solution which requires the minimum investment. It is the solution of the formulation (2.2a). The corresponding first order optimality condition is given by (2.2b).

\[
\min_{y_i} \sum_r C_r(N_r) + \sum_{v_i} B^i v_i
\]

\[B^i \leq L^i \quad v_i \geq 0 \quad N_r \geq 0 \quad z_s, \quad c_s \geq \sum_{v_i} v_i (\frac{-\partial B^i}{\partial N_s}) \forall s, \ s.t. \ N_r > 0 \]

Two new notations are being introduced in (2.2). The first new notation is \(L^i\), which specifies the user desired threshold on the GoS for connections of the OD pair (i,j). The second new notation is \(v_i\), it is the Lagrange multiplier associated with 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 \(v_i\) 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 \(B^j\) is a fundamental component of the formulations F1 and F2. The actual functional form of \(B^j\) varies with routing schemes [4]. We take a different perspective and derive it by using the insight that the blocking function can be approximated by the link connection intensities (at equilibrium) and the link capacities [4]. Techniques from the reliability theory [7] were employed to devise the general functional form for \(B^j\), regardless of the actual routing scheme employed. The \(B^j\) 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 \(R_i\) be a set that contains all these links. Define an indicator variable \(y_i\) for the link \(s\), whereas \(y_i\) equals to zero if link \(s\) has enough resource to admit at least one connection, and \(y_i\) equals to one if link \(s\) does not have resource to serve any connection. The expected value of \(y_i\) 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 \(Y = [y_i]\) as the input. The complement of it, \(\bar{\phi}(Y) = 1 - \phi(Y)\) 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 \(y_i\) are independent zero-one random variables and \(\phi(Y)\) is a Boolean function, we can
perform the Shannon decomposition on the function $\phi(Y)$. By using an arbitrary link $s$ as the pivot we have expression (2.3).

$$
\phi(Y) = y_{s}(1,Y) + (1 - y_{s})\phi(0,Y) \\
= \phi(0,Y) + [\phi(1,Y) - \phi(0,Y)]y_{s}
$$

(2.3)

Where $(0,Y)$ and $(1,Y)$ are the status vectors that differ only in the $s^{th}$ link. The functions $\phi(0,Y)$ and $\phi(1,Y)$ indicate that whether the system has been blocked given that the link $s$ is in admissible status has been blocked. By the definition of $v_{s}$, the expectation $E[y_{s}]$ is the blocking probability 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 $E[y_{s}|y_{s}]$ are replaced by the Erlang-B Loss formula $E_{s}(\cdot)$ and the vector $E_{s}|y_{s}$ respectively in (2.4). The vector $E_{s}|y_{s}$ denotes the collection of Erlang-B loss functions for all the links $s'$ such that $s' \neq s$. The continuous extension of Erlang-B formula suggested in [1] is being used throughout this article and it is shown in (2.5).

$$
E[\phi(Y)] = \int f_{s}(E_{s},\lambda) + \int f_{s}(E_{s},\lambda)E_{s}(\cdot)
$$

(2.4)

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

$$
E_{s}(\lambda, N_{s}) = \left[ \lambda_{s}^{N_{s}} e^{-\lambda_{s}(1 + z^{N_{s}})} \right]^{-1}
$$

(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 $B^{\phi}$s on the (equilibrium) link connection intensities and link capacities.

$$
B^{\phi} = \int f_{s}(\lambda_{s}, N_{s}, y_{s}) + \int f_{s}(\lambda_{s}, N_{s}, y_{s})E_{s}(\lambda_{s}, N_{s})
$$

(2.6)

Expression (2.6) is valid for any link $s$. So we can represent $B^{\phi}$ in the form of $f_{s} + f_{s}E_{s}(\lambda_{s}, N_{s})$ for any link $s$, where $f_{s}$ and $f_{s}$ are independent of the link $s$.

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 Newton’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, $E_{s}(\cdot)$ is the Erlang-B formula as defined in (2.5), $\lambda_{s}$ is the connection intensity on link $s$, $N_{s}$ is the capacity allocated on link $s$. Then the following lemma holds.

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

$$
c_{s} = v\left[ \frac{\partial E_{s}(\lambda_{s}, N_{s})}{\partial N_{s}} \right]
$$

(3.1)

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

**Theorem 1:** Assume the optimization approach in [3] is employed to solve F1 and assume it converges to the optimal solution. Moreover assume the routing scheme does not decrease the (equilibrium) link connection intensity as the capacity of the link increases. Then the optimal GoS derived by F1 is an increasing function of the service charge vector $W=[w]^{T}$.

Proof: Suppose the optimization approach in [3] 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.

$$
c_{s} = \sum_{s} (\lambda_{s}^{w}) \times f_{s}(\lambda_{s}, N_{s}) \left[ \frac{\partial E_{s}(\lambda_{s}, N_{s})}{\partial N_{s}} \right]
$$

(3.2)
Now consider two vectors of service charges, \( W = [w_i] \) and \( W' = [w_i' + \Delta w_i] \), where \( \Delta w_i > 0 \). Assume that the optimal solution with respect to the vector \( W \) is denoted by the tuple \( (\Lambda^W, N^W(\Lambda^W)) \), where \( \Lambda^W = [\lambda_i^W] \) is the link connection intensity vector decided by some optimal routing rules, \( N^W(\Lambda^W) = [N_i] \) 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 \( W' \)). Substitute \( (\Lambda^W, N^W(\Lambda^W)) \) into the optimality condition (3.2), note that the service charges are now \( W' \), and we have
\[
\sum_i (\lambda_i^W + \lambda_i^W \Delta w_i) \times f_i^W(\lambda_{ij}, N_{ij}) > \sum_i (\lambda_i^W \Delta w_i) \times f_i^W(\lambda_{ij}, N_{ij})
\]
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 \( f_i^W \) increases at the end of the first iteration.

This makes
\[
\sum_i (\lambda_i^W + \lambda_i^W \Delta w_i) \times f_i^W(\lambda_{ij}, N_{ij})
\]
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 \( N^W(\Lambda^W) \).

\[ \square \]

Theorem 2: Assume the user-desired GoS guarantees are denoted by \( L_i^{ij} \), (i.e. the users of OD pair \((i,j)\) desires an end-to-end blocking probability of \( B_i^{ij} \) less than \( L_i^{ij} \).) 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 \( L_i^{ij} \) as constraints.

\[
w_i^{ij} = v_{ij} / \lambda_i^{ij} \tag{3.3}
\]

Proof: The minimum cost assignment which satisfies the desired GoS levels satisfy the equations in (2.2b). Assume the operator design the SON network 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 level if the operator designs the network using F1. \( \square \)

Theorems 1 and 2 can be illustrated through the use of a figure. Figure 2 shows the profit contours for a one-link network. There is only one OD pair, \((i,j)\), for this example. The origin (gateway \(i\)) and destination (gateway \(j\)) are connected together by the link. Each blue line in figure 2 corresponds to the expected profit from the link under a particular connection service charge \( w_i^{ij} \). 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 \( w_i^{ij} \), thus the red dots denote the optimal solutions of F1 under the parameters \( w_i^{ij} \). For clear illustration, an arrow is drawn to point the direction of increasing \( w_i^{ij} \) in the objective contour. As \( w_i^{ij} \) 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. Obviously, the x value of this particular red dot is the optimal solution of the formulation F2 with the GoS constraint (since it is the minimum capacity to satisfy the GoS constraint). Yet the same x is also an optimal solution of F1 with respect to a particular service charge value, \( w_i^{ij} \). It turns out the reward \( w_i^{ij} \) on this very red dot corresponds to the value of \( w_i^{ij} = v_{ij} / \lambda_i^{ij} \) 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 \( w_i^{ij} \), 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 that this set of service charges is a Pareto efficient solution to a multi-criteria optimization problem.
**B. Pareto efficient pricing**

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

\[
\begin{align*}
\max_{w^j} f_o &= \max_{w^j} \sum_{i,j}^N \lambda^i w^j (1 - B^j) - \sum_{i,j}^N C_i (N_i) \\
\max_{w^j} f_i &= \sum_{i,j} U_q (B^j, w^j) \\
\text{s.t.} & \quad 0 \leq w^j \leq M^j \quad \forall i,j
\end{align*}
\]  

(3.4)

**Lemma 2:** \(f_o\) is an increasing function of \(w^j\).

**Proof:** Consider the case that \(w^m\) is increased by \(\Delta w^m > 0\) for a particular OD pair \(mn\). Assume the original service charge vector to be \(W=[w^j]\). Denote a new reward vector by \(W'\) where \(W'\) is larger than \(W\) by \(\Delta w^m\) at the \(mn\)th element. Substitute \(W'\) into \(f_o\) at the optimal solution of the original \(W\) (i.e. \(B^j\) and \(N_i\) remain unchanged), the value increases even without re-optimization. Since the value of \(f_o\) can only increase after re-optimization, so we have \(f_o(W') > f_o(W)\). □

It can be shown that the set of service charges defined in (3.3) is a minimum Pareto efficient solution to the multi-criteria 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 \(w^j\) in (3.4) is bounded and closed, the feasible set of (3.4) is compact. Now consider two possible deviations of \(w^j\).

- a) If \(w^j\) is increased by a positive amount of \(\Delta w^j\), 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 \(U_q\) we know that \(U_q(w^j) > U_q(w^j + \Delta w^j)\). So \(f_o\) is improved by increasing \(w^j\) (by Lemma 2) but \(f_i\) is worsened, and the choice of \(W\) is not dominated by the choice of \(W'\).

- b) If \(w^j\) is decreased by a positive amount of \(\Delta w^j\), then according to lemma (2), \(f_i\) decreases. Moreover since \(W' < W\), then according to theorems 1 and 2, the GoS desired by the \(j\)th users is not satisfied. From the definition of user utility we have \(U_q(L', w^j) > U_q(L', M^j) > U_q(L, w^j - \Delta w^j)\). Therefore both \(f_o\) and \(f_i\) are worsened, and the choice of \(W\) is dominated by the choice of \(W'\).

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

**IV. A Simple Example**

Consider a simple SON network in figure 3. Assume there are three Poisson streams of connections, with intensities \(\lambda_{AB}=10\) units per unit time, \(\lambda_{CB}=15\) units per unit time, \(\lambda_{AC}=20\) 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 respectively 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 performance metric of the network and F1 is 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>
<thead>
<tr>
<th>Table 1. Capacity allocation results for low service charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service charges (18.2, 17.8, 18.6)</td>
</tr>
<tr>
<td>Formulation F1</td>
</tr>
<tr>
<td>GoS ((\lambda_{AB}, \lambda_{CB}, \lambda_{AC}))</td>
</tr>
<tr>
<td>(0.084, 0.086, 0.085)</td>
</tr>
<tr>
<td>Allocated capacities on links ((AB, CB, AC))</td>
</tr>
<tr>
<td>(13, 18, 23)</td>
</tr>
<tr>
<td>Cost</td>
</tr>
<tr>
<td>334</td>
</tr>
<tr>
<td>Objective value</td>
</tr>
<tr>
<td>-417.13</td>
</tr>
<tr>
<td>Expected Profit rate</td>
</tr>
<tr>
<td>417.13</td>
</tr>
</tbody>
</table>
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, moreover 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