Optimization Approaches for Service Overlay Network Capacity Allocation

Ngok Lam[1], *Student Member; IEEE*, Zbigniew Dziong[2] and Lorne G. Mason[3], *Senior Members; IEEE*

***Abstract*— We study the capacity allocation problem in service overlay networks (SON)s with state-dependent connection routing based on revenue maximization. The dimensioning problem is formulated as two separate yet related optimization problems, namely the optimal routing problem and the optimal capacity allocation problem. The optimal routing problem is solved by employing an efficient event-dependent routing scheme that approximates the state dependent routing scheme. A call-admission-control (CAC) agent is included to allow reward sensitive connection GoS. This CAC mechanism provides an additional degree of control in the capacity allocation process and further improves the efficacy of the solution scheme. The optimal capacity allocation problem itself is solved by two different approaches; one employs the conventional gradient approach, the other uses the concept of average link shadow price that can be derived from information local to the links. The two approaches together with the CAC agent constitute four different capacity allocation schemes. They are tested and compared with one another. A key contribution of this paper is the extension of the shadow price method to solve a constrained capacity allocation problem efficiently. The second contribution is the illustration of the potential performance gains by allowing an additional degree of control in the optimization process.**

***Index Terms*—Service Overlay Network, Capacity Allocation, Grade of Service, Reward Maximization, Event-Dependent Routing, Optimization methods.**

I. Introduction and Problem Descriptions

The Internet primarily provides best-effort delivery for the data. It does not attempt to differentiate the data nor provide a QoS guarantee to the data. Combining this with the fact that the Internet consists of a collection of independent autonomous systems (ASes), there is a good reason why offering end-to-end quality of service (QoS) guarantees in the pure Internet setting is difficult. Though protocols like DiffServ have been proposed to offer differentiated QoS in the Internet, the presence of independent autonomous systems still makes it difficult to offer end-to-end QoS guarantees across the Internet, as this requires the establishment of multi-lateral business relationships with all the independent ASes the traffic would traverse. It is under that background that Service Overlay Networks, (SON)s, were put forward in the hope of addressing this difficulty. A Service Overlay Network (SON) is a logical network formed on the top of the physical Internet that spans through different autonomous systems. A SON operates in the same way as a Virtual Network, the SON operator owns the SON gateways which are placed in strategic locations. A SON operator leases bandwidths with QoS guarantees from the underlying ASes, usually in the form of Service Level Agreements (SLAs). These bandwidths provide logical connections in the overlay network. Once the bandwidths are in place, the SON is realized and is ready to offer end-to-end QoS guarantees for the value-added services it provides (i.e. VoIP, Video On Demand services, etc). The SON connections are classified by the source and the destination gateways. Users with access to the Internet can access the service gateways and use the value-added services. Figure 1 shows an example of SON network.

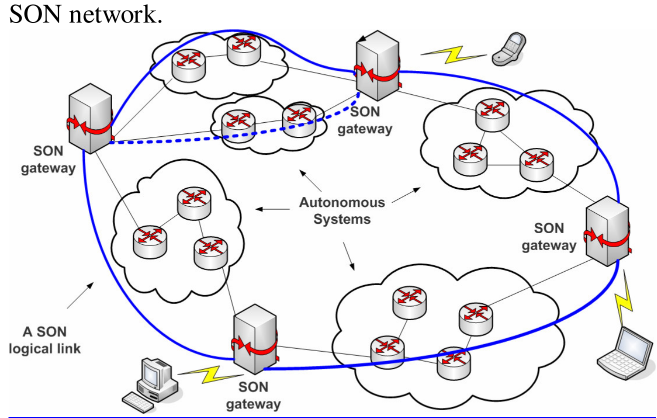

Fig. 1. A SON example.

In order to deploy a SON, the optimal amount of bandwidth to be leased for each of the logical links is a major challenge faced by the SON operator. From the operator point of view, the bandwidth allocated should provide the maximum economic benefit yet meet the user expectation regarding the connection acceptance ratio. This ratio is also known as the Grade of Service requirements (GoS). So the problem confronting the SON operator is: Given a set of SON gateways, what are the optimal bandwidths to lease for the logical links if the traffic intensities between the SON gateway pairs are known and the

Manuscript received 30th, September, 2008. This work was supported in part by grants from NSREC under the project code SP208010.

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

2. Zbigniew Dziong is with the Department of Electrical Engineering, École de technologie supérieure, Montreal, Quebec, Canada H3C 1K3 (email: zdziong@ele.etsmtl.ca)

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

GoS requirements are given. This is also the question we address in this article. To facilitate the study, we shall consider the SON networks as generic networks, and the SON gateways as generic networking nodes. The objective is to solve the optimal bandwidth allocation problem of the generic networks. We assume all the traffic follows the exponential distribution for the inter-arrival time and the service time. The traffic demands are assumed to be homogenous(the optimization framework can be extended to accommodate heterogeneous traffic in a relative straight forward manner). They consist of the service specific data flows between the source-destination SON gateways. It is also assumed that the traffic revenue rates are given parameters. We also assumed that traffic among the source-destination node pairs are routed by state-dependent routing schemes such as the MDPD scheme proposed in [1]. We first solve the capacity allocation problem in an economic framework by incorporating a *novel routing scheme* into the optimization formulation as in [2] to approximate the state-dependent routing scheme assumed in SON. Then we derive a capacity allocation scheme based on the notion of link shadow prices in the routing layer. Moreover, the same shadow prices are employed to approximate the CAC functionalities of a network so that CAC functionality is accounted for in the optimization formulations.

Traditionally the literature for the network dimensioning problem is related to circuit switched networks such as the telephone network. Some of the works in the literature that are related to ours are [3], [4], [5], [6], [7], [8], [9], [10], [11] and [12]. These studies usually assumed fixed routes [7], [8], [9], [12] or some simple routing strategies [3] ,[4], [6], [10], [11]. The problem is usually formulated as an optimization problem that optimizes certain objective [3], [4], [5], [7], [8], [9], [10], [11], [12]. Instead of employing the profit function as the objective to be maximized, many authors [7], [8], [10], [11] chose to employ the cost function as an objective to be minimized. While the minimum cost formulation may fit well some of the problems, it is not necessarily the same problem that the SON operators are interested in solving. Among the abovementioned works, [9] and [12] are direct studies of the SON capacity allocation problem. In article [9] Duan et al studied the capacity allocation problem without having hard GoS constraints. The capacity allocation problem was formulated and solved approximately. By assuming the Guassian distribution for the link traffic, Park et al [12] refined the results of [9] and showed that problem formulated in [9] can be approximated by a separable convex optimization problem, Mitra et al have a series of papers on capacity allocation for the virtual networks [13], [14]. Their approach is closest to the one we are taking. They derived two approximations for the sensitivities of the blocking probabilities with respect to the capacities [15], [16]. By using these sensitivity approximations, Mitra was able to approximate the average link shadow price using Kelly's average shadow price concept [17]. In [13], [14] a network design problem based on the aforementioned approximation of average link shadow price was formulated. The optimization criterion is the network reward instead of network profit (i.e. cost ignored). Mitra's model differs from the model we study here. First we are employing an event-dependent routing strategy that is driven by the blocking of the connections in the model. Second there is the addition of the GoS constraints in our model. Third the objective of our optimization approach is the expected profit rate of the SON network, which is the performance metric of interest under the SON environment.

The article is structured as follows: section 2 will be devoted to the discussion of the concepts of link shadow price and the path shadow price. An event-dependent routing algorithm for approximating the state-dependent routing schemes assumed in the SON environment is described in section 3. The mathematical formulations of the optimization models are included in section 4. Numerical and simulation results are presented in section 5. Section 6 concludes the study.

II. Link Shadow Prices and Path Shadow Prices

The concept of link shadow price has existed in the routing literature since the 1980s. In [17] Kelly introduced the concept of implicit link cost due to accepting a connection in circuit-switched networks. The average link shadow price (ALSP) discussed in [17] is independent of the link status. It indicates, on an average, the loss in network revenue by removing one unit of bandwidth from a link. State dependent link shadow price (SDLSP) on the other hand, is an indication of the expected loss in revenue by removing one unit of bandwidth from a link in a given link state. ALSP can be derived from the SDLSP by taking the expectation of SDLSP with respect to the network status or by considering the left hand side derivative of the profit function with respect to the link capacity [1], [17]. The calculations of SDLSP and ALSP only require information local to the links. It is particularly convenient for decentralized processing within a network. The concept of link shadow price (both ALSP and SDLSP) has been widely employed in the routing literature as a means to measure the profitability of accepting a connection [1], [18]. Shadow price concepts are also used in network Call Admission Control (CAC) of various routing schemes to maximize the profit from the routing layer. The SDLSP indicates the expected loss by removing one unit of bandwidth from the link in a particular link state. It can act as an indication of the implicit cost in using the link. It can also act as an indication of the profitability of a link.

We assume the MDPD approach for routing and CAC decisions in a SON. When a connection arrives at a SON gateway, the CAC module of the SON checks the reward and calculates the so-called path net gains for the connection on all the candidate paths, the definition of path net gain for a path is given in expression (1). Where w^{ij} is the reward from the connection, r is a candidate path for the connection, and $p_s(x_s)$ is the shadow price of the link s at state x_s[1]. The equation describes the expected gain on the candidate path by accepting the connection.

$$w^{ij} - \sum_{s \in r} p_s(x_s) \quad (1)$$

The connection will be admitted to a path with the maximum non-negative path net gain as given by expression (2). We assumed the connection will be admitted if and only if there is a positive net gain in the path.

$$\arg\max_{r \in R_{ij}} \{ w^{ij} - \sum_{s \in r} p_s(x_s) \} \quad (2)$$

Using the statistical link independence assumption, the network reward process can be decomposed into separable link processes as in [18]. The SDLSP of a link can be derived conveniently using the first passage time of the M/M/c/c Markov chain at each of the link states. The resulting expression for SDLSP of homogenous traffic is a recursion as stated in (3), and can be efficiently calculated with a time complexity of $O(N)$, where N is the link capacity. We assume all connection arrivals follow the Poisson distribution. The Poisson connection arrival rate to the link s from OD pair ij is λ_s^{ij} in equation (3), N_s is the capacity of the link, w_s^{ij} is the reward assigned to link s by the OD pair ij and $\overline{w_s}$ is average reward rate of offered traffic.

$$p_s(n) = \frac{\lambda_N}{1-\mu_N r_{N-1}} \frac{\lambda_{N-1}}{1-\mu_{N-1} r_{N-2}} \cdots \frac{\lambda_{n+1}}{1-\mu_{n+1} r_{n-1}} \overline{R}_s \quad (3a)$$

$where:$

$$\lambda = \sum_{ij} \lambda_s^{ij}, \lambda_i = \frac{\lambda}{\lambda + i \cdot \mu}, \mu_i = \frac{(i-1)\mu}{\lambda + i \cdot \mu}, r_i = \frac{\lambda_{i+1}}{1-\mu_{i+1} r_{i-1}}, r_0 = \lambda_1 \quad (3b)$$

$$\overline{w_s} = \frac{\sum_{ij} \lambda_s^{ij} w_s^{ij}}{\sum_{ij} \lambda_s^{ij}} \quad (3c)$$

$$n = 0, \ldots, N_s - 1$$

A candidate path for a connection can consist of multiple links. The information from the path shadow price distribution can be employed to facilitate a form of Call Admission Control for the routing algorithm such that only connections with positive net gain will be admitted by the path. To derive the shadow price distribution of a multi-link path, one needs to perform multiple convolutions with the SDLSPs. The straight forward convolution method would yield a time complexity of $O(N^n)$ in deriving the path shadow price distribution, where $N = \max_{i \in r}\{N_i\}$, N_i is the link capacity of link i, and $n=|r|$ is the number of links contained in this path r. Due to the high time complexity, this method can not be effectively applied to facilitate analytical studies. We put forward two methods to alleviate this difficulty. The first method involves applying a two-moment approximation for the distribution of the path shadow price. Again, by assuming link independence, the mean and variance of the path shadow price is approximated by the sum of the corresponding link means and variances. The path mean and variance are the input parameters of a carefully chosen distribution to approximate the real path shadow price distribution. This method has the particularly simple complexity of $O(nN)$ where $N = \max_{i \in r}\{N_i\}$ and n is the number of links in the path. A major deficiency of this approach is that the accuracy of approximation is not always guaranteed. In general it is doubtful whether it will be easy to choose a distribution that would match well with the convoluted shadow price under different network conditions. The path shadow price distribution is usually multi-modal with multiple peaks at the modes of the individual link shadow prices, and its shape changes according to the link loads. Moreover, since the path shadow price is a discrete distribution, the real path shadow price distribution is somewhat "peaky" as shown in the figure 3. Nevertheless when no other tools are available, the particular simple two-moment approximation is still worthwhile to give a try.

The second method we proposed is an approximation method based on aggregation. Similar ideas in simplifying the convolution calculation by quantization and aggregation can also be found in [1] and [19]. The SDLSP defined in (3a) is the expected loss of revenue due to the removal of one bandwidth unit from the link. From (3a-3c) it can be seen that the SDLSP is always bounded by the average revenue of its offered traffic, as the probability of reaching the blocking state from any states is at most 1. As a consequence the convoluted path shadow price is also bounded by the sum of its links' average revenues. Define by $UB_{r(i)}$ the upper bound for the convoluted shadow price of the first i links of the route r; instead of performing the convolution blindly, the shadow prices can be aggregated into the equal-partitioned intervals [*0, delta*), [*delta,2*delta*), [*(k-1)*delta*,$UB_{r(i)}$] after performing the $(i-1)^{th}$ convolution. There are k intervals after each convolution. We define the error for the approximated path shadow price to be the largest possible deviation of the approximated shadow price value from its real value. Since $UB_{r(i)}$ is non-decreasing in i, the maximum possible error due to quantization for this convolution methodology is bounded by the following expression $\varepsilon = \sum_{i=1}^{n-1} \frac{UB_{r(i)}}{2k} \le (n-1)\frac{UB_{r(n-1)}}{2k}$. Note that $UB_{r(i)} > UB_{r(j)}$ if $i>j$. The real shadow price value lies surely in the interval $(x_a - \varepsilon, x_a + \varepsilon)$, where x_a is the approximated path shadow price value. It can be shown that if k is of O(n), this error can be made arbitrarily small. Even if the k is of a moderate magnitude, it can be observed that errors of distribution functions are still reasonably small. An example of approximated cumulative function is shown in figure 2 below, which compares the distribution functions for the real distribution and the two approximations. The time complexity of this convolution scheme is $O(N_1^* N_2^* + n^2 N)$, where N_1^* and N_2^* are the capacities of the two smallest-capacity links in the path. N is the capacity of the largest-capacity link, n is the number of hops in the path. In practice the term $n^2 N$ dominates the time complexity for paths with large number of links.

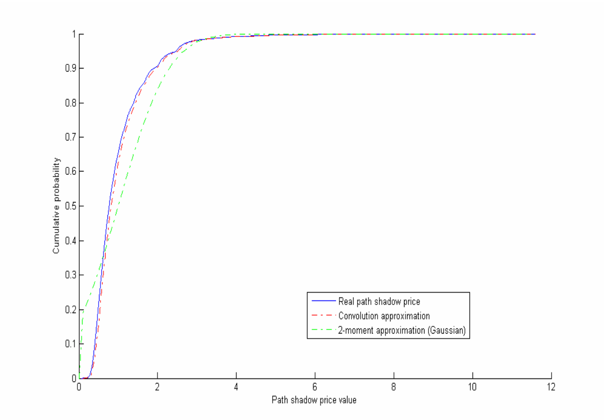

Fig. 2. Cumulative distribution of a path's shadow price

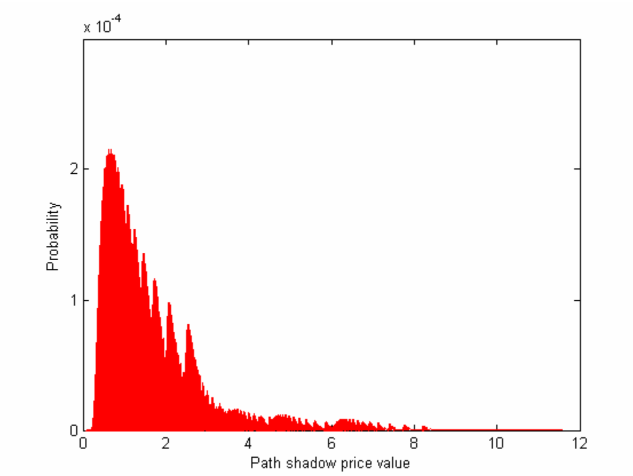

Fig. 3. Probability distribution of a path's shadow price

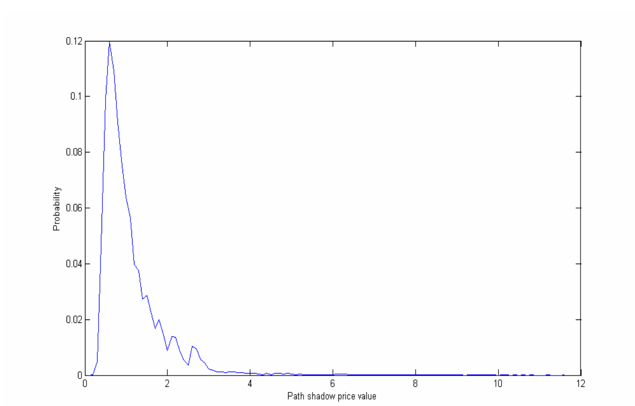

Fig. 4. Probability distribution of a path's shadow price (approximation)

The time taken to get the distribution in figure 3 is over 8 minutes on a P4 1.7Ghz machine, yet the time taken to get the distribution in figure 4 is 0.2 second on the same P4 machine. The detailed pseudo-code for the algorithm is listed in table 1. This approximation allows us to employ the path shadow price to enforce a form of CAC in the analytical model efficiently. In particular we can use the cumulative distribution function of the path shadow price to set up a CAC "filter" that filters out traffic which does not offer enough reward to use the path (i.e. those give negative path net gain). This is also the way that the CAC module of state-dependent routing like the MDPD scheme would handle the traffic. The addition of this "filter" attempts to capture the characteristic of the CAC module. The "filter" enables the traffic GoS performance to be sensitive to the traffic reward, which effectively differentiates the traffic with respect to their reward parameters, and offers an additional degree of control over the resource in the capacity allocation process.

Table 1 : A fast convolution Algorithm

pick the two smallest-capacity links 1^* and 2^* within the path *r*
for i=1 to N_{1^*}
  for j=1 to N_{2^*}
    perform convolution on the links 1^* and 2^*
  end
end

for i=1 to $N_{1^*} \times N_{2^*}$
  1.allocated arrays
    SP_K([*0, delta*), [*delta,2*delta*), [*k*delta,*$UB_{r(2)}$]) and
    $PROB_K$([*0, delta*), [*delta,2*delta*), [*k*delta,*$UB_{r(2)}$])
  2.aggregate the convoluted shadow price and probability into the corresponding locations in SP_K and $PROB_K$
end

for i=2 to n-1
  for j=1 to $N_{i\in r} \neq N_1^*, N_2^*$
    convolute SP_k, $PROB_k$ with SP_i and $PROB_i$
  end

  1.allocated arrays
    SP_K([*0, delta*), [*delta,2*delta*), [*k*delta,*$UB_{r(2)}$]) and
    $PROB_K$([*0, delta*), [*delta,2*delta*), [*k*delta,*$UB_{r(i+1)}$])
  2.aggregate the convoluted shadow price and probability into the corresponding locations in SP_K and $PROB_K$
end

return SP_K and $PROB_K$

III. Routing Algorithm

To incorporate the MDPD state-dependent routing algorithm into the optimization model is computationally infeasible, as that will require solving a complete Markov decision problem in every iteration of the capacity allocation process. In order to approximate the state-dependent routing algorithm assumed in the SON environment we propose an event-dependent routing scheme which is a modified version of the second routing scheme in [2]. In this proposed routing scheme, each of the potential paths for a particular traffic flow is assigned a load sharing coefficient [20]. These load sharing coefficients can be calculated according to the path net gain distributions (ie. proportion of the time that the path has the maximum path net-gain among all the potential paths for the traffic flow. This method will be elaborated subsequently). These load sharing coefficients can alternatively be treated as optimizing variables and calculated from an optimization model. The load sharing coefficients are directly proportional to the probabilities that the traffic will be routed through the path. In the case when there are *n* possible paths for a particular traffic, the scheme will select a path among the n possible paths for the traffic flow to attempt. If this path is blocked, the scheme will attempt the remaining $n-1$ paths with probabilities proportional to the paths' load sharing coefficients. If the new path chosen by the scheme also turns out to be blocked, the session request will attempt the remaining $n-2$ paths with probabilities proportional to their original load sharing coefficients. This process continues until either the traffic flow for the session is routed or until the set up process discovers that all n paths are blocked. We assume this routing scheme keeps a record of the previously attempted blocked paths, so that they will be avoided as the algorithm proceeds. Figure 5 depicts a case of such a routing scheme. In this scenario, path *P1* is discovered blocked by a session attempt assigned to it. The traffic therefore overflows to the remaining paths *P2* and *P3*, with probabilities directly proportional to their load sharing coefficients a_2 and a_3.

The major difference between the routing scheme employed in this paper and the 2nd routing scheme in [2] is that in this new scheme, blocking can occur not only when there is insufficient bandwidth in the links, but may also occur when the reward provided by the traffic is not high enough to cover the implicit cost of using the path (this is dependent on whether the "filter" is present). Unlike routing scheme *II* in [2] the new routing scheme uses only local information and each flow records the blocking information itself (no exchange of data at the source node). So for the same link blocking pattern in the network,

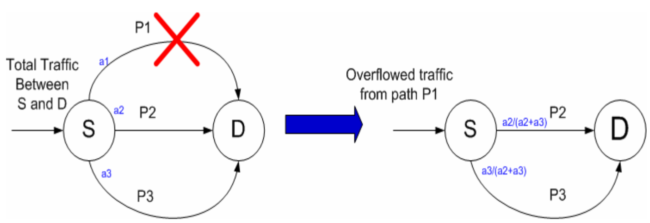

Fig. 5. An illustration of the routing algorithm

there are several possible sequences for a session to select the possible paths. The optimization formulations are therefore different from that of [2] and they are listed in section 4. It is worthwhile to mention that the optimal sets of load sharing coefficients give system optimality for routing inside the network, and provide the maximum profit to the SON operator under this event-dependent routing scheme. The load sharing coefficients may be considered as an approximation of the proportion of time that the paths are having the highest path net gain in the MDPD routing scheme. Combining this load sharing coefficient with the "filter" that calculates the probability such that the net gains are non-negative, this event-dependent routing scheme effectively gives a reasonable approximation to the MDPD routing scheme we assumed in the SON environment. Incorporating the filter into the routing scheme also provides an additional degree of freedom in allocating network resources in the optimization models. We shall see in section 5 that the incorporation of the filter is able to deliver further network performance gains in the optimization process.

IV. THE MODELS

In order to solve the capacity allocation problem, two sets of decision variables need to be calculated, namely the load sharing coefficients and then the link capacities. In other words, the optimal routing problem and the optimal link capacity problem need to be solved together as the solutions of the routing problem affects the capacity allocation problem and vice versa. We shall propose two approaches to solve the problem, and then we shall compare the results from these approaches in sections 5 and 6.

The first optimization approach follows the same line as discussed in [2]. We shall present a brief introduction to it. In this approach the capacity allocation problem is formulated as a maximum profit optimization problem. We shall denote it by the notation "OPT" in the remainder of this article. We solve the routing problem and the dimensioning problem both using the optimization framework. The key idea is to solve for the decision variables that correspond to the set of KKT conditions derived. In the following description, we assume N_s, α_r^{ij}, B_s are the independent variables. N_s is the capacity of the link s. α_r^{ij} is the load sharing coefficient (or routing coefficient) that denotes the probability that the network would route the traffic of the OD pair ij (denoted by f_{ij}) through path r. $\Theta_{r,k}^{ij}$ is a set containing all the *k-permutations* of the candidate paths (excluding the path r) for the traffic ij. $o_{r,k}^{ij}$ is a set containing a particular *k-permutation* (also excluding path r), and $o_{r,k}^{ij}(h)$ is an overloading of the expression to indicate another set that contains only the first h paths in the set $o_{r,k}^{ij}$. The variables N_s and α_r^{ij} are the decision variables to be solved. B_s is a dummy variable to alleviate part of the complexity involved in deriving the first order derivatives with respect to N_s. The optimization formulation, the Lagrange function and the corresponding KKT equations are given below in (5), (6) and (7a-c). The objective function in (5) has two parts. The first part, $\Sigma C_s(N_s)$, donates the total cost rates incurred from allocating the capacities in the network. The second part represents the total expected reward rates from the individual routes. Condition (7a) corresponds to the set of optimal routing equations, note the multiplier v^{ij} is independent of the route r, by taking into account the complementary slackness condition, condition (7a) implies all routes carrying positive flow must have the same first derivative path length. Condition (7c) corresponds to the condition that marginal reward of each of the links equals to its marginal cost at optimality. The key component of (7c) are the y multipliers, which are derived from condition (7b). The capacity allocation problem (5) is not necessarily convex nor uni-modal, as confirmed by our numerical study. Multiple runs with different initial solutions could be required to guarantee the quality of the solution. A simple neighborhood check procedure may be carried out to ensure the local optimality of the solution if the second order check is to be avoided. An interesting problem that guarantees further investigation is whether there exists instances of the problem such that the problem is convex.

In the case that the CAC "filter" is incorporated, one needs to modify the definition of blocking probability to take into account the addition blockings due to the filter, as well as the thinning of the offered traffic due to the presence of the filter. Moreover, owing to the presence of the "filter" in the routing layer, a penalty approach in the routing layer may be required to avoid the filter from overriding the GoS requirement. Whenever required the penalty term can be calculated in a sub-gradient-like approach as shown below in (4):

$$p_{ij}^{(i+1)} = [p_{ij}^{(i)} + (\bar{L}_{ij} - L_{ij}^{i})s_i]^{+} \tag{4}$$

where $p_{ij}^{(i+1)}$ is the penalty term for traffic pair ij at the (i+1)[th] iteration, L_{ij}^{i} is the GoS level for traffic pair ij at the ith iteration, $\bar{L}_{ij}$ is the required GoS level, s_i is the step size at the i[th] iteration.

$$\min \sum_{s} C_s(N_s) - \sum_{ij}\sum_{r\in R_{ij}} \{\alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r) + \alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r)[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})]\}$$

$s.t.$

$$1-\prod_{r\in R_{ij}} B_t \geq \bar{L}_{ij} \qquad x^{ij}$$

$$\sum_{r\in R_{ij}} \alpha_r^{ij} = 1 \qquad v^{ij}$$

$$\alpha_r^{ij} \geq 0 \qquad u_r^{ij}$$

$$E(a_s, N_s) = B_s \qquad y_s$$

$$N_s \geq 0 \qquad z_s \tag{5}$$

$$\begin{aligned} L = {} & \sum_{s} C_s(N_s) - \sum_{ij}\sum_{r\in R_{ij}} \alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r) \\ & -\sum_{ij}\sum_{r\in R_{ij}} \alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r)[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})] \\ & +\sum_{ij} x^{ij}(\prod_{r\in R_{ij}} B_r - \bar{L}_{ij}) + \sum_{ij} v^{ij}(\sum_{r\in R_{ij}} \alpha_r^{ij} - 1) - \sum_{ij}\sum_{r\in R_{ij}} \alpha_r^{ij}u_r^{ij} \\ & +\sum_{s} y_s(E(a_s,N_s) - B_s) - \sum_{s} z_s N_s \end{aligned} \tag{6}$$

$$\begin{aligned} v^{ij} = {} & u_r^{ij} + \lambda^{ij}w_r^{ij}(1-B_r) + \\ & \lambda^{ij}w_r^{ij}(1-B_r)[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})] \\ & + \sum_{r'\in R_{ij}} \alpha_{r'}^{ij}\lambda^{ij}w_{r'}^{ij}(1-B_{r'}) \frac{\partial[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})]}{\partial\alpha_r^{ij}} \\ & -\sum_{s} y_s \frac{\partial E(a_s,N_s)}{\partial\alpha_r^{ij}} \end{aligned} \tag{7a}$$

$$\begin{aligned} y_s = {} & \sum_{ij}\sum_{\substack{s\in r\\ r\in R_{ij}}} \alpha_r^{ij}\lambda^{ij}w_r^{ij}\frac{\partial B_r}{\partial B_s} \\ & +\sum_{ij}\sum_{\substack{s\in r\\ r\in R_{ij}}} \alpha_r^{ij}\lambda^{ij}w_r^{ij}\frac{\partial B_r}{\partial B_s}[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})] \\ & -\sum_{ij}\sum_{s\in r} \alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r)\frac{\partial[\sum_{k=1}^{|R_{ij}|-1}\sum_{b_r^{ij}\in\Theta_k^{ij}}\sum_{O_k^{ij}}((\prod_{\substack{h=1\\ r'\in R_{ij}\\ r'\notin O_k^{ij}(h)}}^{k}\frac{1}{\sum\alpha_{r'}})(\prod_{r''} B_{r''}\alpha_{r''})]}{\partial B_s} \\ & +\frac{\partial\sum_{ij} x^{ij}(\prod_{r\in R_{ij}} B_r)}{\partial B_s} + \sum_{\substack{s'\\ s'\neq s}} y_{s'}(\frac{\partial E(a_{s'},N_{s'})}{\partial B_s}) \end{aligned} \tag{7b}$$

$$C_s = -y_s\frac{\partial E(a_s,N_s)}{\partial N_s} + z_s \tag{7c}$$

The second approach is an approximation scheme. It involves solving the optimal routing problem, and then deriving the link shadow prices from the optimal routing policy utilizing knowledge local to the links. Based on the shadow prices, the optimal link capacities can be obtained. We shall denote it by the notation "SP" in the remainder of this article. The optimal routing problem formulation is similar to that of the previous approach, and is listed below in (8), (9) and (10). The same penalty approach may also be required when the filter is present. The objective in (8) represents the total expected reward rates from the individual routes r of the OD pair ij. Condition (10) corresponds to the optimal routing equation. By taking into account the complementary slackness condition, condition (10) again implies all routes carrying positive flow must have the same first derivative path length.

$$\min -\sum_{ij}\sum_{r\in R_{ij}} \{\alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r) + \alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r)[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})]\}$$

$s.t.$

$$\sum_{r\in R_{ij}} \alpha_r^{ij} = 1 \qquad v^{ij}$$

$$\alpha_r^{ij} \geq 0 \qquad u_r^{ij} \tag{8}$$

$$\begin{aligned} L = {} & -\sum_{ij}\sum_{r\in R_{ij}} \alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r) \\ & -\sum_{ij}\sum_{r\in R_{ij}} \alpha_r^{ij}\lambda^{ij}w_r^{ij}(1-B_r)[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})] \\ & +\sum_{ij} v^{ij}(\sum_{r\in R_{ij}} \alpha_r^{ij} - 1) - \sum_{ij}\sum_{r\in R_{ij}} \alpha_r^{ij}u_r^{ij} \end{aligned} \tag{9}$$

$$\begin{aligned} v^{ij} = {} & u_r^{ij} + \lambda^{ij}w_r^{ij}(1-B_r) + \lambda^{ij}w_r^{ij}(1-B_r)[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})] \\ & + \sum_{r'\in R_{ij}} \alpha_{r'}^{ij}\lambda^{ij}w_{r'}^{ij}(1-B_{r'}) \frac{\partial[\sum_{k=1}^{|R_{ij}|-1}\sum_{O_{r,k}^{ij}\in\Theta_{r,k}^{ij}}((\prod_{h=1}^{k}\frac{1}{\sum_{\substack{r'\in R_{ij}\\ r'\notin O_{r,k}^{ij}(h)}}\alpha_{r'}^{ij}})(\prod_{\substack{r''\\ r''\in R_{ij}\\ r''\in O_{r,k}^{ij}(h)}} B_{r''}\alpha_{r''}^{ij})]}{\partial\alpha_r^{ij}} \end{aligned} \tag{10}$$

In this formulation the optimal routing problem is first solved, the traffic intensities on the links are then known. By applying equations (3a-3c) and taking the expectations, the average link shadow prices can be derived readily for different link capacities. Average link shadow price at a given link capacity can be viewed as a marginal revenue generated from the last link [1]. The link cost on the other hand is the marginal cost of the total link cost. Maximum profit is achieved when the average link shadow price equals to the link cost, this can be viewed as a correspondence of the first order optimality condition (of the maximum profit formulation). To tackle the GoS constraints the optimal capacity problem is solved based on the following augmented set of equations

$$c_s = R_s(\lambda,\alpha,\overline{w}_s,N_s)+\sum_{ij\in\theta_s} x^{ij}\frac{\partial B_{ij}}{\partial N_s}$$
$$= R_s^A(\lambda,\alpha,\overline{w}_s,N_s,x^{ij}) \qquad (11)$$

Where c_s is the cost of the link, λ is the average traffic intensity vector, α is the load sharing coefficient vector, $\overline{w}_s$ is average reward rate of link s as defined in (3c), N_s is the capacity of link s, and function R_s maps the parameters λ, α, $\overline{w}_s$ and N_s to the average shadow price of link s according to expressions (3a-3c). θ_s is a set containing the indexes of all OD pairs that have link s in their routes. Since c_s, λ, α, and $\overline{w}_s$ are given parameters, x^{ij} is the set of multipliers which can be derived through sub-gradient method, the value of N_s such that equality symbol holds in (11) can be calculated. Assume continuous extension of Erlang B equation [21] is being employed, $R_s^A(.)$ is a strictly decreasing in N_s, and there is an one-to-one correspondence between the function's value and N_s. Therefore (11) can be conveniently solved by either Newton's method or by linear search methods like bisection search. In theorem 1 of the appendix, it was shown that the formulation based on expression (11) is approximately equivalent to the maximum profit optimization formulation discussed earlier, they are not exactly equivalent owing to simplifications made in decomposing the network reward process into the link reward processes. Nevertheless it provides a simple approximation to the derivatives in expression (7). Because of its simpler performance model, this approach delivers significant performance gain over the full gradient approach, as we shall see in section 5. To derive the link shadow price, the connection rewards on a path w_r^{ij}, are required to be divided into w_s^{ij} on the links of the path. Various link decomposition rules had been reported in the literature, see for example [1], [18]. In our investigation, it was found that equal division of reward (i.e. connection reward divided equally to all the links constituting the path) gives the best solutions among the division rules and therefore this rule is used throughout the study described in this article. The economic interpretation for this division scheme is, since every link offers the same amount of resource in carrying a connection, the link should be offered the same amount of reward for carrying the connection as any other links in the path.

V. Numerical and Stimulation Results

Numerical studies were performed on a fictitious network example as shown in figure 6. The outgoing traffic intensities are directly proportional to the population of the source node,

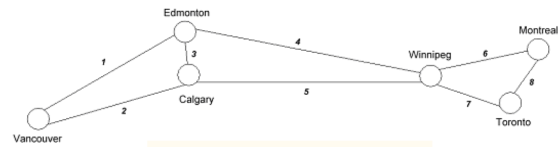

Fig. 6. A fictitious Canada SON network

and are equal to the cities' population divided by 20 000, the outgoing traffic are assigned to the destination cities according to the ratio of their population to the total city population. The GoS constraints are set to 2% for all the OD pairs (i.e the minimum service level is 98%). The traffic matrix is listed in table 2 below as a reference. The unit of traffic intensity in the table is Erlang. The link costs are proportional to its physical distances and are listed in table 3. The rewards are random numbers listed in the table 4. For the sake of generality we do not force the reward matrix to be symmetric.

Table 2: traffic rate matrix

To From	Vancouver	Edmonton	Calgary	Winnipeg	Montreal	Toronto	Total Outflow
Vancouver	X	3.000	3.900	2.600	4.300	11.200	25
Edmonton	3.125	x	5.625	3.750	6.250	16.250	35
Calgary	4.150	5.850	X	5.000	8.350	21.650	45
Winnipeg	2.650	3.700	4.750	X	5.250	13.700	30
Montreal	4.700	6.600	8.500	5.650	X	24.550	50
Toronto	17.550	24.600	31.600	21.100	35.150	X	130
Total Inflow	32.175	43.750	54.375	38.100	59.300	87.350	315

Table 3: link cost rates per unit bandwidth

Link index	1	2	3	4	5	6	7	8
cost	5	4	2	7	7	11	9	3

Table 4: connection reward rate matrix

To From	Vancouver	Edmonton	Calgary	Winnipeg	Montreal	Toronto
Vancouver	x	12	14	24	35	20
Edmonton	25	X	21	17	30	18
Calgary	19	31	X	23	27	30
Winnipeg	31	22	16	X	21	31
Montreal	27	18	20	24	X	35
Toronto	19	25	35	32	15	X

We performed the capacity allocation for four different approaches; the first two approaches use the optimization framework discussed in section 4. We apply the traffic filter and combine it with the routing mechanism to one approach (denoted by OPT_Filter). While the other one uses the same routing mechanism yet it does not include the filter (denoted by OPT_NonFilter). The remaining two approaches are based on the local link information encapsulated by the link average shadow price, we include the filter in one (SP_Filter) and do not include it in another (SP_NonFilter). Network simulator based on the MDPD approach [1], [18] is employed to verify GoS requirements. Confidence interval information regarding profit rates is also gathered from the same simulator. Table 5 gives the capacity allocation of the four methods.

Table 5:

Link Index	1	2	3	4	5	6	7	8	Profit rate (Analytical)	Profit rate (simulation)
OPT Filter	26	53	26	94	129	68	174	91	3414.2	3422.8±101
OPT NonFilter	20	55	25	92	133	67	173	100	3416.9	3426.9±112
SP Filter	14	56	44	69	154	38	203	114	3401.6	3437.0±112
SP NonFilter	17	52	49	72	152	39	199	115	3410.5	3448.7±103
OPT Pure Load Sharing	17	52	67	107	133	87	158	143	2975.4	3041.8±102

For illustration, a plot of the average profit rates is shown in the figure 7 below. The blue dots correspond to simulation averages, the black vertical lines donate the 95% confidence intervals, and the pink dots are the average profit rates calculated from the analytic performance model. The time durations of the four capacity allocation schemes on a P4 2.0Ghz machine are also shown in the figure 7 as blue bars, note the logarithmic scaled Y-axis. It can be seen that owing to the simpler performance model the shadow price approaches provide similar network profit with roughly half of the time durations required by the gradient approaches.

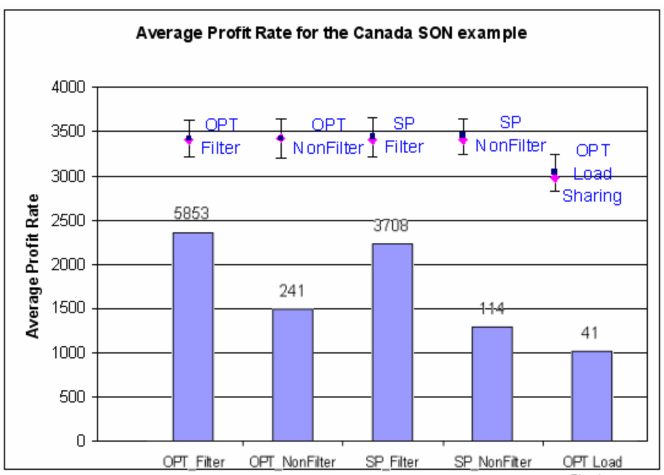

Fig. 7. The Canada SON network example, the average profit rate and the time durations.

For quick verification, the analytical and simulation GoS values of the 30 traffic pairs are also plotted in figures 8a and 8b. All the GoS values are below 0.02 which is the service level being set in this example. From the capacity assigned in table 5, we can see the major difference between the SP and the OPT approaches is that the SP approaches assign more capacities to links that have high traffic intensities (i.e links 5 and 7 that are present in 12 and 15 of all the 34 active paths respectively) and low link costs, yet they also tend to assign less capacities to links that has lower traffic intensities than what the OPT approaches would do (i.e. links 1, 4 and 6). This phenomenon is observed in other examples too. The reason is that the SP approaches are some hill-climbing optimization procedures based on local knowledge (i.e. knowledge such as link traffic intensities, link costs and link average rewards). Local traffic intensities and link costs have a profound effect on the hill-climbing directions. Therefore links with high traffic

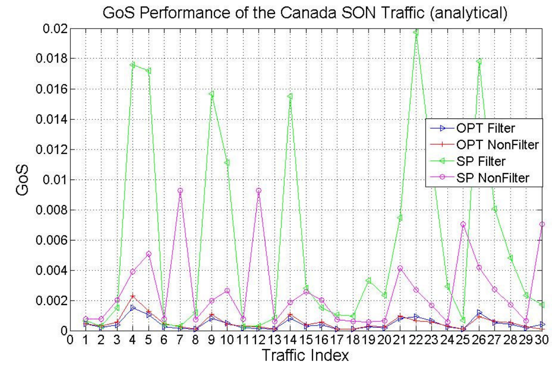

Fig. 8a. GoS performance of the Canada SON traffic (analytical)

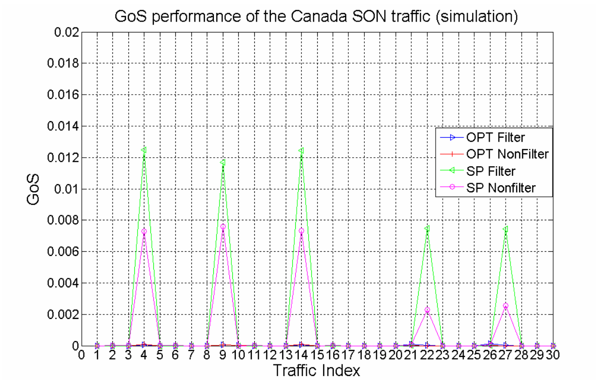

Fig. 8b. GoS performance of the Canada SON traffic (simulation)

intensities and low cost would appear to the SP approaches as the better candidates for resource allocation. Moreover it was also observed that the SP approaches tend to assign larger capacities to links within a minimum spanning tree (MST) of the graph. As we have mentioned, the link costs has a significant impact on the improvement directions. When link costs are low, the SP approaches are likely to allocate the links with large capacities, so the augmentation of capacities usually occurs over the "safe edges" [20] of the network (i.e. the cheapest links). The way that link capacities get augmented is similar to the way that the Kruskal's algorithm behaves in finding a minimum spanning tree by choosing the "safe edges". When the "safe edges" selected by the SP approaches join together, they form a MST(provided that no cycles are formed). This augmentation of capacities along the MST is evident in the example above, over links 3, 5 ,7 and 8. Whereas these four links form a minimum spanning tree (MST) connecting the nodes E, C, W, M and T.

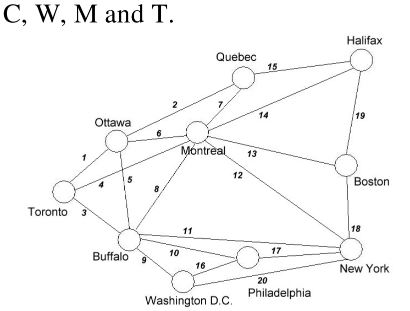

Fig. 9 A fictitious North America SON network

Another sample run was performed on a larger fictitious North America SON example in figure 9 with 20 links, 45 OD pairs and 135 routes.The results are shown in the figure 10 below, the blue bars correspond to the running times. For the sake of simplicity only the profit rates and the run time results are shown. Note that this example is roughly two times larger than the previous Canadian SON example, there are 20 links (vs 8 links) and 135 routes (vs 60 routes). It can be seen in figure 10 that all three approaches based on the traditional gradient approach (including OPT load sharing) required roughly quadratic (400%-600%) more running time for this network. Yet the SP approaches, in particular the "nonfilter" approach, requires only approximately linear (210%) more running time. The "filter" version of the SP approach suffers from the time penalty due to filter calculations for all the 135 routes, yet it requires less than 50% of the time duration required by the OPT approach while giving a solution in a similar quality to the OPT filter approach. This shows the scalability benefit by having a simpler performance model.

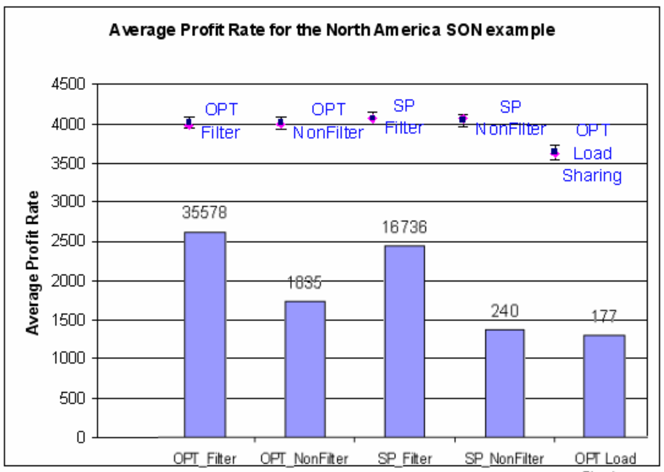

Fig. 10. The North America SON network example, the average profit rate and the time durations.

It is worthwhile to mention that although the "filtered" approaches in these two fictitious SON examples do not offer performance gain in terms of network profit rates, the ability to differentiate traffic streams with respect to their reward, however, provides a valuable degree of control in the network. As an illustrative example, we shall inject five new traffic pairs VE, EW, CW, WM and WT respectively to the fictitious Canadian SON example. Assume the newly added traffic pairs having traffic connection rates of 30 units, reward rates of 2 units and they all require a GoS of 0.3. The benefit of adding a filter to enable an additional degree of control becomes evident as shown in the figure 11.

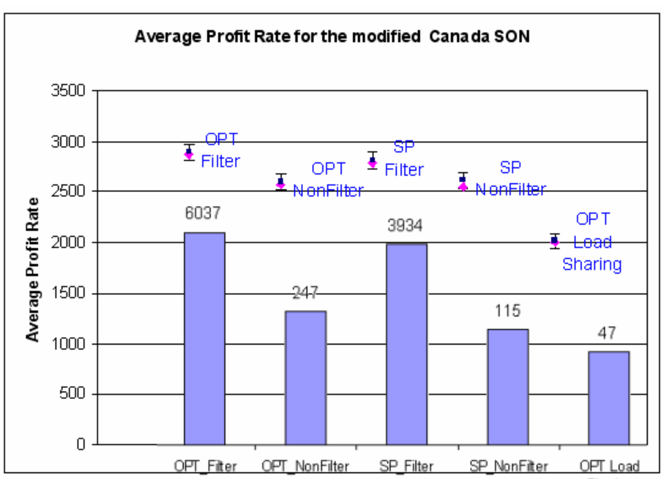

Fig. 11. The modified Canada SON network example, the average profit rate and the time durations.

As can be seen, in order to accommodate the newly injected traffic, the profit rates now decrease. This is due to the fact that the injected connections' rewards are so low that they can not even cover their fees for using the links. Note that in this modified case the filter approaches are able to deliver significant better performance in terms of average profit rate, the gain from having a filter is roughly 10% in both OPT and SP approaches. To facilitate the explanation of this phenomenon, we resort to a simple one-link example as shown in figure 12.

Fig. 12. A one-link network example

In this example we assume there are two classes of traffic from node A to node B. Index the first traffic by 1 and the second by 2. We have the following information, $\lambda^1=6$ $w^1=1500$ $GoS^1=0.02$, $\lambda^2=5$ $w^2=150$ $GoS^2=.30$, link cost for link A-B is 250 per unit bandwidth per unit time. Using the pure optimization approach without the filter we get the optimal capacity $N_{AB}=18$, and profit rate at 5106.0, $GoS_1= GoS_2= 0.0148$. Employing the optimization approach with filter, however, we get the optimal capacity $N_{AB}=16$, profit rate 5380.8, $GoS_1= 0.019$, $GoS_2=0.2642$. The filtered approach is able to offer better performance by allowing an extra degree of control over the connection admission (even though the granularity of the filter may not be fine in this small example). Network resources can be utilized more reasonably without offering unnecessarily high GoS service levels. When GoS levels are similar and reward rates are close as in the Canadian and North America SON examples, the filtered approaches are likely to offer solutions with similar quality as the non-filter approaches. In particular when the GoS levels are similar, there is little room for the filter to enforce GoS differentiation. But when the network traffic are highly differentiated in nature with large gaps in rewards and GoS requirements, and when network resources are expensive, embedding a CAC agent in the capacity allocation process would deliver solutions with significantly better quality. For networks with rather uniform traffic rewards and GoS requirements, the non-filtered approaches are preferable because of the lower computational overhead. Otherwise the filtered approaches can be employed to take advantage of the added degree of control in resource utilization.

VI. Conclusions and Future Works

We have performed a comprehensive investigation of the Service Overlay Network dimensioning problem and provided some novel ingredients to solve the problem. To be specific, we provided a framework for the SON dimensioning problem by two different approaches, one based on the tradition gradient method, the other based on the information provided by the average link shadow price. We showed that these two approaches are approximately equivalent. Because of the simpler performance model, the shadow price approach involves significantly less computational burden than the traditional gradient approach. We compared the quality of the solution and we realized they are similar in terms of network profit generation. We also derived a traffic filter based on link shadow price distribution to further refine the control over the network resource utilization. It was shown that though the addition of the extra degree of control imposes higher computational overhead, this extra degree of control is valuable in improving the solution quality for networks with highly differentiated traffic. We are now investigating the way to use a hill-climbing approach in getting the optimal filtering probability with respect to the GoS constraints. This effort is expected to significantly reduce the computational overhead in incorporating the GoS differentiation mechanism in the optimization process. Moreover we are also looking for further refinement and simplification to the link shadow price calculation. With these future refinements, we expect our new methodology to be able to handle large networks yet still offer a reasonably fine degree of control over the resource utilization so as to suit a large variety of traffic classes with different reward rates and GoS requirements.

Appendix

Theorem 1:
The shadow price capacity allocation approach is approximately equivalent to the optimization approach in [2].

Proof:
Hiding the routing part and abstracting it by B_{ij}, the blocking of traffic for OD pair ij, the optimization formulation of [2] is equivalent to (T1), the notation employed here are the same as those mentioned earlier in this report.

$$\begin{aligned} \min \quad & \sum_s C(N_s) - \sum_{ij} \lambda^{ij} w^{ij} (1 - B_{ij}) \\ & B_{ij} \le \bar{L}^{ij} && x_{ij} \\ & E(a_s, N_s) = B_s && y_s \\ & N_s \ge 0 && Z_s \end{aligned} \tag{T1}$$

The set of KKT conditions are shown in (T2) and (T3), where $c_s>0$ is the cost of allocating one unit of capacity on link s.

$$\frac{\partial L}{\partial N_s} = c_s + \frac{\partial E(a_s, N_s)}{\partial N_s} y_s - Z_s \tag{T2}$$

$$\frac{\partial L}{\partial B_s} = \sum_{ij} \lambda^{ij} w^{ij} \frac{\partial B_{ij}}{\partial B_s} + \sum_{ij} x_{ij} \frac{\partial B_{ij}}{\partial B_s} - y_s \tag{T3}$$

Substituting (T3) into (T2) we have (T4):

$$\begin{aligned} c_s &= -\frac{\partial E(a_s, N_s)}{\partial N_s} y_s + Z_s \\ &= -\frac{\partial B_s}{\partial N_s} \Big(\sum_{ij} \lambda^{ij} w^{ij} \frac{\partial B_{ij}}{\partial B_s}\Big) - \frac{\partial B_s}{\partial N_s} \Big(\sum_{ij} x_{ij} \frac{\partial B_{ij}}{\partial B_s}\Big) \\ &= -\Big(\sum_{ij} \lambda^{ij} w^{ij} \frac{\partial B_{ij}}{\partial N_s}\Big) - \Big(\sum_{ij} x_{ij} \frac{\partial B_{ij}}{\partial N_s}\Big) \\ &= \frac{\partial R}{\partial N_s} - \sum_{ij} x_{ij} \frac{\partial B_{ij}}{\partial N_s} \\ &\approx R_s(\lambda, \alpha, \bar{w}_s, N_s) - \sum_{ij} x_{ij} \frac{\partial B_{ij}}{\partial N_s} \\ &= R_s^A(\lambda, \alpha, \bar{w}_s, N_s, x_{ij}) \end{aligned} \tag{T4}$$

Assume the reward function R is a differentiable function of the capacity N_s (i.e. the continuous version of Erlang equation is employed). The term $\frac{\partial R}{\partial N_s}$ is the right hand side derivative of the total reward w.r.t the capacity N_s. A change in N_s may bring changes to the offered traffic to links other than s, which in turn changes reward rates on these links. It is in fact a derivative of the *network reward* w.r.t. N_s, as it measures the changes in network reward owning to the change of N_s. The term $R_s^A(\lambda, \alpha, \bar{w}_s, N_s)$ calculated based on (3a-3c) and (11), is a left hand side approximation of the derivative of the *link reward*. When a change of capacity in link s brings negligible reward changes in other parts of the network, the two expressions are equivalent, and if this is the case the two capacity allocation processes are also equivalent.

Acknowledgment

The authors would like to thank NSERC for their funding support for Strategic Project SP208010**.**

References

[1] Z. Dziong, "ATM network resource management", McGraw-Hill, 1997.

[2] N. Lam, Z. Dziong and L.G. Mason, "Network capacity allocation in service overlay networks", in *Proc. 20th International Teletraffic Congress (ITC)*, Ottawa, 2007, pp. 224-235.

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

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

[5] S. Shi, J. S. Turner, "Multicast routing and bandwidth dimensioning in overlay networks", *IEEE Journal on Selected Areas of Communication*, Vol 20, No. 8, pp. 1444-1455, 2002,.

[6] A. Girard and B. Sanso, “Multicommodity flow models, failure propagation, and reliable loss network design”, *IEEE transactions on Networking*, Vol. 6, No. 1, pp. 82-93, 1998.
[7] D. Medhi, D. Tipper, “Some approaches to solving a multi-hour broadband network capacity design problem with single-path routing”, *Telecommunication Systems* ,Vol. 13, pp. 269–291, 2000.
[8] B. Gavish, I. Neuman, “A system for routing and capacity assignment in computer communication networks”, *IEEE Transactions on Communications*, Vol. 34, No. 4, pp. 360-366, 1989.
[9] Z. Duan, Z. L. Zhang, and Y. T. Hou, “Service Overlay Networks: SLAs, QoS, and bandwidth provisioning”, *IEEE/ACM Transactions on networking*, Vol.11, No. 6, pp. 870-883, 2003.
[10] B. Sanso, A. Girard, and F. Mobiot, "Integrating reliability and quality of service in networks with switched virtual circuits", *Computers and Operations Research*, Vol. 32, pp. 35-58, 2005.
[11] D.C. Dietz, A.J. Elcan, and D.E Skipper, "Optimization Models for ATM network planning", *Computer and Operations Research*, Vol. 30, pp. 625-641, 2003.
[12] K.J. Park and C.H. Choi, “Optimization driven bandwidth provisioning in service overlay networks”, Computer Communications, Vol 31, pp. 3169-3177, 2008
[13] D. Mitra, J. A. Morrison and K. G. Ramakrishnan, “ATM Network Design and Optimization: A Multirate Loss Network Framework”, *IEEE/ACM Trans. on Networking*, Vol.4, No.4, pp. 531-543, 1996.
[14] D. Mitra, J.A. Morrison and K.G. Ramakrishnan, “Virtual Private Networks: Joint Resource Allocation and Routing Design”, in *Proc. IEEE INFOCOM 99*, New York, 1999, pp. 480-490.
[15] D. Mitra and J. A. Morrison, “Erlang capacity of a shared resource”, in *Proc. 14th International Teletraffic Congress(ITC)*, Kluwer, 1994, pp. 875-885.
[16] J. A. Morrison, K. G. Ramakrishnan and D. Mitra, “Refined asymptotic approximations to loss probabilities and their sensitivities in shared unbuffered resources”, *SIAM Journal Appl. Math.* , Vol. 59, Issue 2, pp. 494-513, 1998.
[17] F.P. Kelly, “Routing in circuit-switched networks: Optimization, Shadow Prices and Decentralization”, *Advanced in applied probability*, Vol. 20, No. 1, pp. 112-144 , 1988.
[18] Z. Dziong and L. G. Mason , “Call admission and routing in multi-service loss networks”, *IEEE transactions on communications*, Vol. 42, No.2/3/4, pp. 2011-2022, 1994.
[19] P.Y Simard, L. Bottou, P. Haffner and Y. LeCun, “Boxlets: a Fast Convolution Algorithm for Signal Processing and Neural Networks”, in *Proc. Advances in neural information processing systems,* Vol. 11, 1999, pp. 571-577.
[20] A. Balakrishnan, T.L. Magnati, J.S. Sokol, and Y. Wang, "Spare capacity assignment for line restoration using a single facility type.", *Operations Research*, Vol 50, pp. 617-653, 2002.
[21] A. Jagers and E. V. Doom “On the continued Erlang loss function”, *Operations Research Letters*, Vol. 5, No. 1, pp. 43-46, 1986.
[22] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, “Introduction to algorithms”, 2nd edition, The MIT Press, 2001.
[23] M. Pioro, D. Medhi, “Routing flow and capacity design in communication and computer networks”, Morgan Kaufmann, 2004.
[24] R. F. Farmer and I. Kaufman, “On the numerical evaluation of some basic traffic formulae”, *Networks*, Vol 8, pp. 153–186, 1978.
[25] D. Bertsekas, R. Gallager, “Data networks”, 2nd edition, Prentice-Hall, 1987.
[26] S. M. Ross, “Introduction to probability models”, 8th edition, Academic Press, 2002.
[27] A. Girard, “Routing and dimensioning in circuit-switched networks”, Addison-Wesley, 1990.
[28] D. Medhi and S. Guptan, ”Network dimensioning and performance of multi-service, multi-rate Loss networks with dynamic routing”, *IEEE transactions on networking*, Vol. 5, No. 6, pp. 944-957, 1997.
[29] M. Liu, J. S. Baras “Fixed point approximation for multi-rate multi-hop loss networks with state-dependent routing”, *IEEE transactions on networking*, Vol. 12, No.2, pp. 361-374, 2004,.
[30] S. P. Chung and K. W. Ross, “Reduced load approximations for multi-rate loss networks”, *IEEE transactions on Communications*, Vol. 41, No. 8, pp. 1222-1231, 1993.
[31] S. I. A. Shah, A. Girard, “Multi-Service network design: a decomposition approach”, in *Proc. Globecom 98*, 1998, pp. 3080-3085.
[32] A. G. Greenberg and R. Srikant “Computational techniques for accurate performance evaluation of multi-rate, multi-hop communication networks”, *IEEE transactions on Networking*, Vol. 5, No. 2, pp. 266-277, 1997.
[33] S.Z. Qiao and L.Y. Qiao, “A robust and efficient algorithm for evaluating Erlang B formula”, Department of Computing and Software, McMaster University, Technical Report CAS98-03, 1998.
[34] D.Mitra, J. A. Morrison and K.G.Ramakrishnan, “Optimization and Design of Network Routing using Refined Asymptotic Approximations”, *Performance Evaluation*, vol. 36-37, No. 1-4, pp.267-288, 1999.
[35] R. F. Serfozo, “Equitable transit charges for multi-administration telecommunications networks”, *Queueing Systems*, Vol. 2, Issue 1, pp. 83-92, 1987.
[36] J. S. Esteves, J. Craveirinha, and D. M. Cardoso, “Second order conditions on the overflow traffic from the Erlang-B system”, Cadernos de Matemática, Universidade de Aveiro, CM06/I-20, 2006.
[37] D. P. Bertsekas, “Nonlinear Programming”, 2nd edition, Athena Scientific, 1999.
[38] C.C. Lo and B.W. Chuang, "An adaptive survivability admission control mechanism using backup VP's for self healing ATM networks", in *Proceedings of the 7th International Symposium on Computers and Communications*, 2002, pp. 647-652.